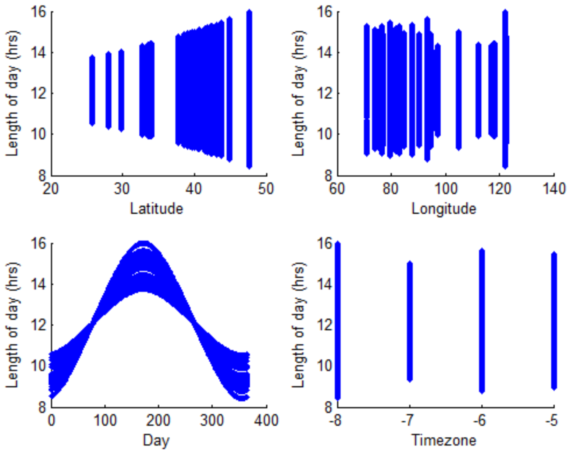
In order to see the effects of the different factors on the length of day, we made the following scatterplots.



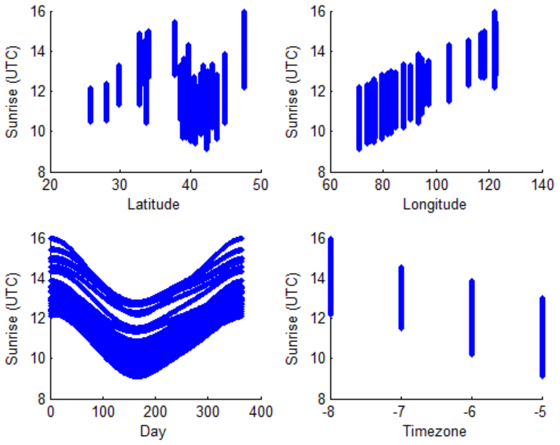
From these scatter plots, we conjecture that length of day seems to be a sinusoidal function of the day with nodes at about day 80 and 12 hrs. The variation of the length of day also seems to increase linearly with latitude, suggesting that it would be a linear multiple of the sine function of day. The length of the variation in the length of day increases from about 3 hrs to 7.5 hrs as latitude increased from about 26 to 48 degrees, so the linear multiple of the sine component might look like:

$$\frac{1}{2}\left(\frac{4.5}{22}Lat-2\right)\sin\left(\frac{day-80}{365}\,2\pi\right)$$

The timezone and longitude does not seem to affect the length of day. So overall:

$$Length\ of\ day \approx 12+\frac{1}{2}\left(\frac{3.5}{22}Lat-2\right)\sin\left(\frac{day-80}{365}\,2\pi\right)$$

For the sunrise, we first convert the local times to UTC time so that the times will be in the same timezone. Then, we plot the scatterplot for sunrise (UTC) vs various factors:



The times seem to vary as a cosine function of the day, with offset of zero since it peaks around day 1 and 365 (period seems to be around 365 days, amplitude seems to be around 1.5 hrs). Latitude does not seem to play a key role here. The times also seem to vary linearly with longitude and timezone, and the longitude seems to capture this linear relationship at a finer level. It seems to increase linearly with longitude with parameters $Slope \approx \frac{14-11}{120-70} = \frac{3}{50}$, $Intercept \approx 6$

$$Sunrise - timezone \approx 6 + \frac{3}{50} Long + 1.5\cos\left(\frac{day}{365}\ 2\pi\right)$$

Putting this altogether:

$$Sunrise - timezone \approx 6 + \frac{3}{50} Long + 1.5\cos\left(\frac{day}{365}\ 2\pi\right)$$

$$Sunset - Sunrise \approx 12 + \frac{1}{2}\left(\frac{3.5}{22} Lat - 2\right)\sin\left(\frac{day - 80}{365}\ 2\pi\right)$$

We can replace all the approximated constants above with parameters A through E such that:

$$Sunrise \approx timezone + A + B\ Long + C\cos(D\ day)$$

$$Sunset - Sunrise \approx E + (F\ Lat - G)\sin(H\ day - I)$$

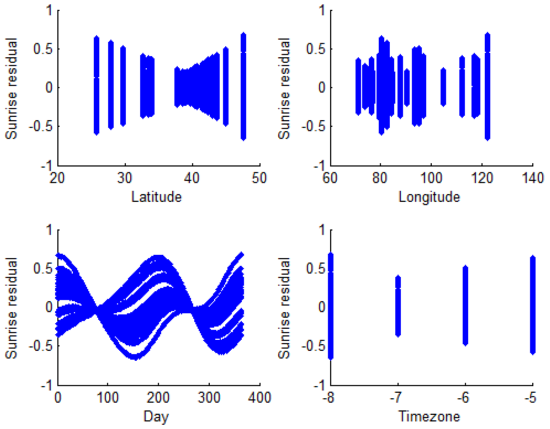
With initial values:

$$A = 6, B = \frac{3}{50}, C = 1.5, D = \frac{2\pi}{365}, E = 12, F = \frac{3.5}{44}, G = 1, H = \frac{2\pi}{365}, I = \frac{160\pi}{365}$$

We turn this over to LSQNONLIN and get the following optimized parameters:

$$A = 5.9, B = 0.07, C = 1.3, D = 0.02, E = 12.1, F = 0.09, G = 0.96, H = 0.02, I = 1.3$$

Clearly, the optimized parameters are quite close to the initial estimates. However, the max norm of the residual in sunrise and length of day was 41 and 19 mins respectively. To gain more insight, we plot the scatterplot of residuals vs factors:



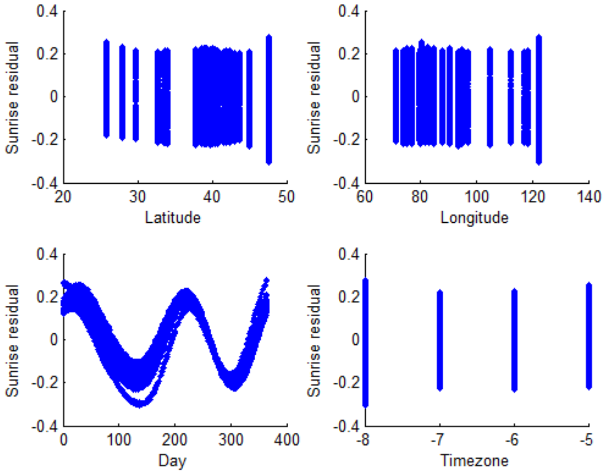
We see that latitude seems to have a multiplicative effect on the residual, and it seems to be periodic, so we can approximate this with a cosine function, and it's effect is likely to be small, so we can try the new functional form of:

$$Sunrise \approx timezone + A + B\ Long + (C - \cos(J\ Lat))\ (\cos(D\ day)\,, \qquad J \approx 0.01\ (small)$$

With this, LSQNONLIN found the following (and other parameters were unchanged since we did not change the other function):

$$A = 5.9, B = 0.07, C = 1.3, D = 0.02, J = 0.05$$

Now, the max norm of residuals for sunrise times is reduced to 18 mins, a good improvement. The residual plot now becomes:



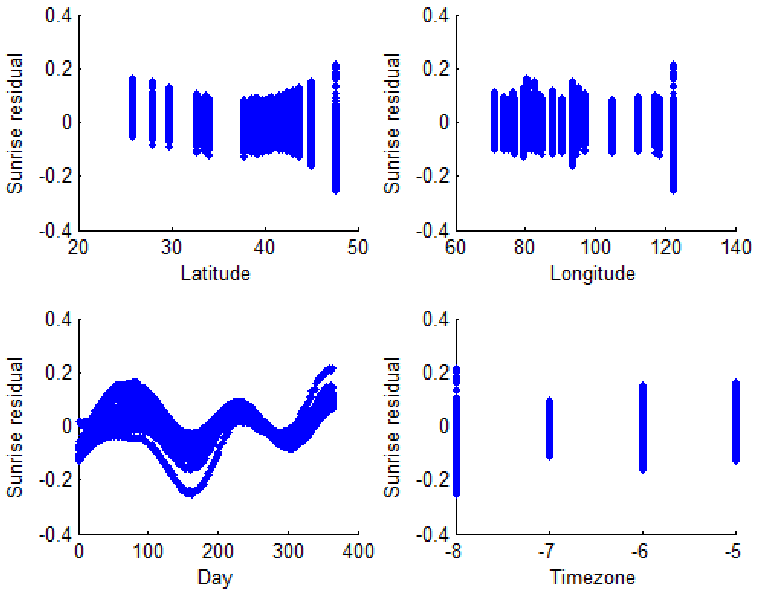
Now, we can clearly see a second cosine function for the day with amplitude about 0.2 and period of about 200 days. Now we modify the function to be:

$$Sunrise \approx timezone + A + B\ Long + (C - \cos(J\ Lat))\ (\cos(D\ day) + K\cos(L\ day)\,,$$
$$K \approx 0.2, L \approx \frac{2\pi}{200}$$

The optimized parameters are:

$$A = 5.9, B = 0.07, C = 1.2, D = 0.02, J = 0.05, K = 0.18, L = 0.03$$

Optimizing, the new scatter plot of residual becomes:



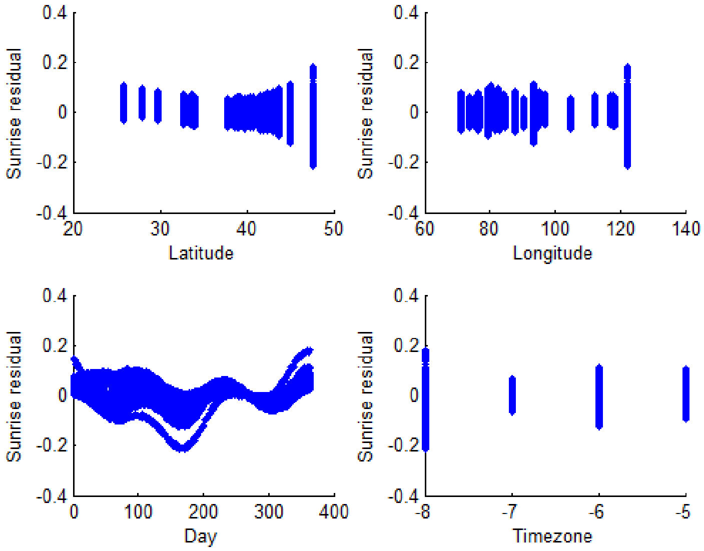
We can still see a sinusoidal effect of day on the residual, so we add in a sine function with amplitude of 0.1 and period of 200 days:

$$Sunrise \approx timezone + A + B\ Long + (C - \cos(J\ Lat))\,(\cos(D\ day) + K\cos(L\ day)) + M\sin(N\ day)\,, \qquad M \approx 0.1, N \approx \frac{2\pi}{200}$$

The optimal parameters are:

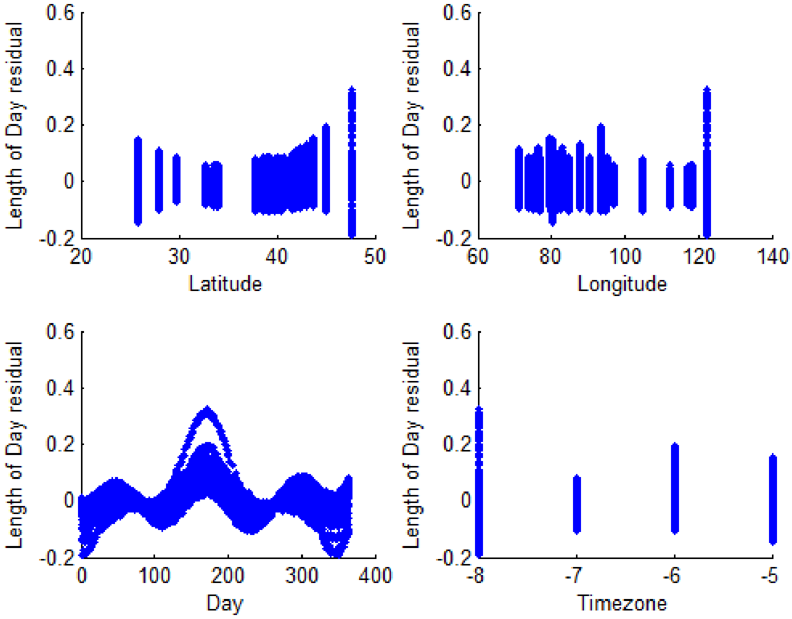
$$A = 5.9, B = 0.07, C = 1.1, D = 0.02, J = 0.05, K = 0.19, L = 0.03, M = 0.1, N = 0.036$$

The residual becomes:



At this point, our function has become rather complicated, and while we still see some high frequency waves in the residual of the day, we would rather not increase the complexity of our model in fear of overfitting our data. The max norm error for sunrise time is 13 mins, and max norm error for sunset time is 8 mins.

The residual scatterplot for length of day is:



We are happy with the low error in length of day fitting, so we don't want to increase the complexity of the model as we might be overfitting, although there is a clear high-frequency sinusoidal function of day present in the residual. . In summary, our fitted function is:

$$
\begin{aligned}
Sunrise &\approx Timezone + 5.9 + 0.07 \times Long + (1.1 + \cos(0.05 \times Lat)) \times \cos(0.02 \times Day) \\
&\qquad + 0.19\cos(0.03 \times Day) + 0.1\sin(0.036 \times Day) \\
Sunset &\approx Sunrise + 12.1 + (0.09 \times Lat - 0.96)\sin(0.02 \times Day + 1.3)
\end{aligned}
$$

Again, the max norm error for sunrise and sunset times are 13 mins and 8 mins respectively. We could potentially improve this by introducing more terms, we could also end up overfitting the data. However, we did not manage to fit our function to within 1 min of the given data.

Overall, LSQNONLIN took 90 and 4 iterations for optimizing sunrise and length of day respectively. This is remarkably fast since our initial estimates were close to the optimal solution found.