Time Series and Crime

Owen Ward

Time Series and Crime

```
library("knitr")
library("tidyverse")
theme_set(theme_minimal())
```

Is crime predictable?

- France created the first centralized system of crime reporting in 1825.
- Guerry (1833) analyzed more than thirty thousand property crimes and ten thousand personal crimes committed between 1825 and 1830.
- The incidence of (reported) crime varied considerably across France. However, regular patterns emerged in the data.
  e.g. crimes against persons consistently highest in summer, crimes against property consistently highest in winter.
- Guerry wondered whether immutable laws—like those describing the phenomena observed in physics—determined crime, ultimately concluding:

  "... the facts of the moral order, like those of the physical order, obey invariant laws, and that, in many respects, the judicial statistics render this a virtual certainty."

- Andre-Michel Guerry (1802-1866) was famous in his lifetime, winning the Montyon Prize twice. But he is largely unappreciated today.
- Friendly (2007) believes Guerry's modesty—both in birth and personality—allowed others to claim credit for his discoveries.
- Nevertheless, his work (along with that of Quetelet) founded the field of "moral statistics" and ultimately sociology and criminology.
- Additional accomplishments: invented the polar/rose plot, invented a mechanical calculator to compare trends, and was mayor of his village.

Lets first construct the data Guery analysed.

```
# personal crimes
tibble(Year    = 1825:1830,
       North   = c(25, 24, 23, 26, 25, 24),
       South   = c(28, 26, 22, 23, 25, 23),
       East    = c(17, 21, 19, 20, 19, 19),
       West    = c(18, 16, 21, 17, 17, 16),
       Central = c(12, 13, 15, 14, 14, 18)) %>%
kable()
```

Year	North	South	East	West	Central
1825	25	28	17	18	12
1826	24	26	21	16	13
1827	23	22	19	21	15
1828	26	23	20	17	14
1829	25	25	19	17	14
1830	24	23	19	16	18

```r
# property crimes
tibble(Year    = 1825:1830,
       North   = c(41, 42, 42, 43, 44, 44),
       South   = c(12, 11, 11, 12, 12, 11),
       East    = c(18, 16, 17, 16, 14, 15),
       West    = c(17, 19, 19, 17, 17, 17),
       Central = c(12, 12, 11, 12, 13, 13)) %>%
kable()
```

Year	North	South	East	West	Central
1825	41	12	18	17	12
1826	42	11	16	19	12
1827	42	11	17	19	11
1828	43	12	16	17	12
1829	44	12	14	17	13
1830	44	11	15	17	13

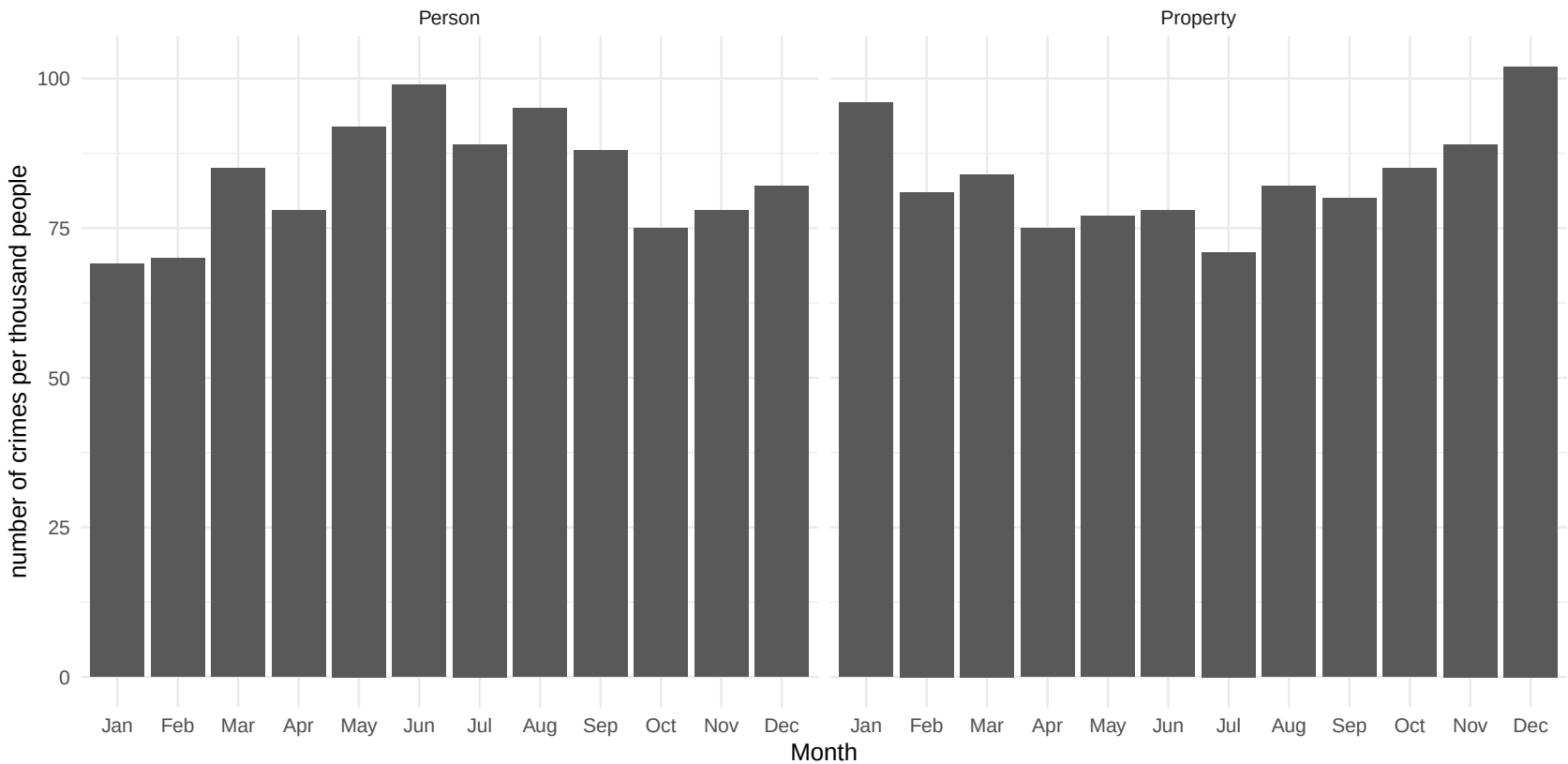
Next we can put these together in a dataset that we will analyse.

```
Guerry <-
 tibble(Month    =
          factor(format(ISOdate(1833,1:12,1),"%b"),
           levels = format(ISOdate(1833,1:12,1),"%b")),
        Person   = c(69, 70, 85, 78, 92, 99,
                     89, 95, 88, 75, 78, 82),
        Property = c(96, 81, 84, 75, 77, 78,
                     71, 82, 80, 85, 89, 102))
Guerry %>%
  top_n(4) %>%
  kable()
```

Month	Person	Property
Jan	69	96
Oct	75	85
Nov	78	89
Dec	82	102

It appears that Person crimes are greater in summer, while property crimes are greater in winter.

```
(guerry_plot <-
Guerry %>% gather(type, rate, -Month) %>%
  ggplot(aes(x = Month, weight = rate)) +
    geom_bar() + facet_wrap(~ type) +
    labs(y = "number of crimes per thousand people"))
```



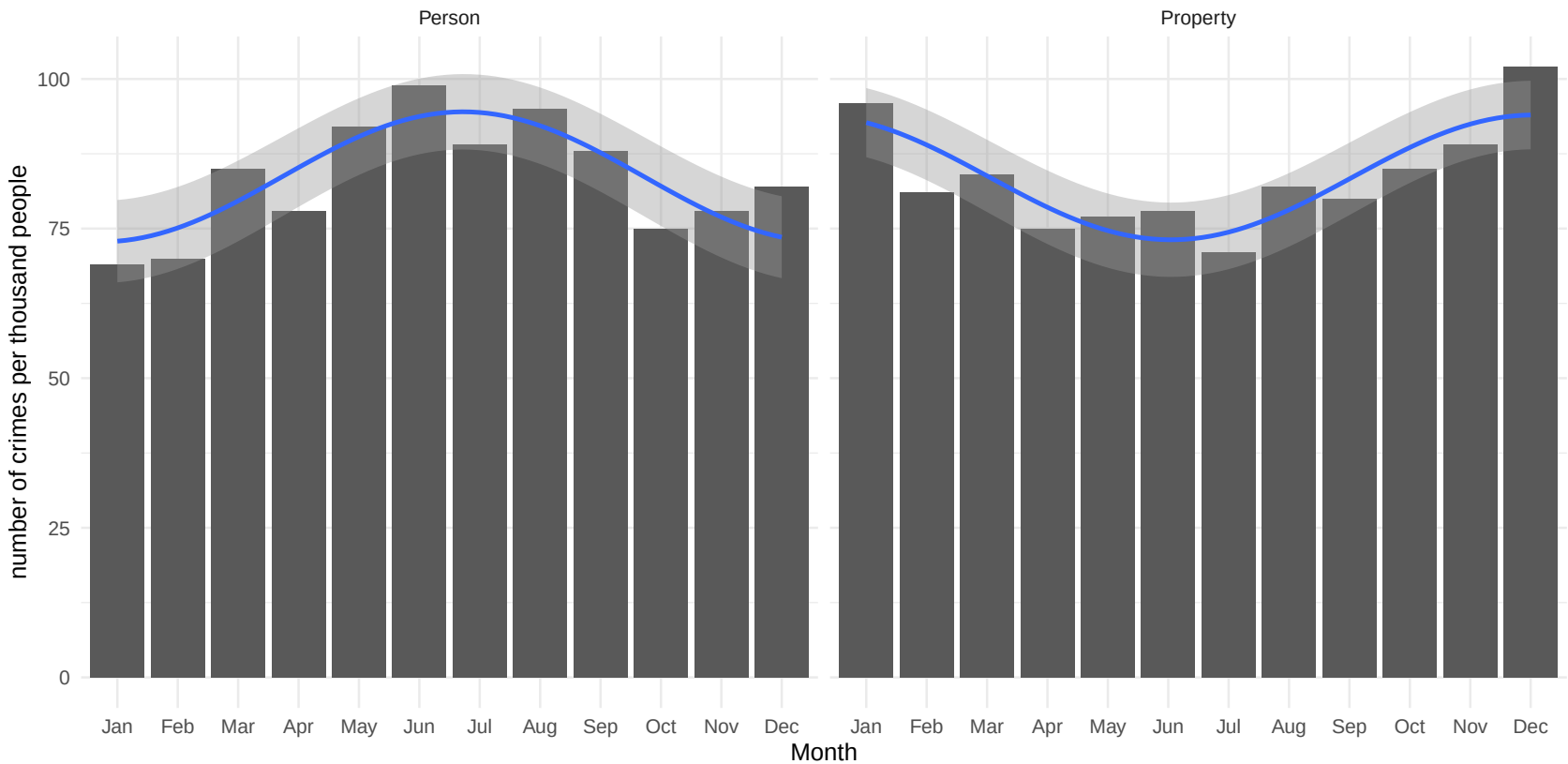
In fact, the seasonal pattern is well approximated by a sine curve.

```
guerry_fit <- Guerry %>%
  gather(type, rate, -Month) %>%
  filter(type == "Person") %>%
  mutate(x = as.numeric(Month)) %>%
  lm(rate ~ cos(x*2*pi/12) + sin(x*2*pi/12),
     data = .) %>% coef() %>% unname()
tibble(
 "$alpha_2$" = guerry_fit[2],
 "$alpha_3$" = guerry_fit[3],
 "$fi$" = atan(guerry_fit[3]/guerry_fit[2]),
 "$A$" = sqrt(guerry_fit[2]^2 + guerry_fit[3]^2)) %>%
  kable(digits = 2)
```

$alpha_2$	$alpha_3$	fi	A
-10.07	-4.18	0.39	10.91

We can look at how this curve looks overlaid on the true data.

```
guerry_plot +
  geom_smooth(aes(as.numeric(Month), rate),
    method = "lm",
    formula = y ~ cos(x*2*pi/12) + sin(x*2*pi/12),
    data = Guerry %>% gather(type, rate, -Month))
```

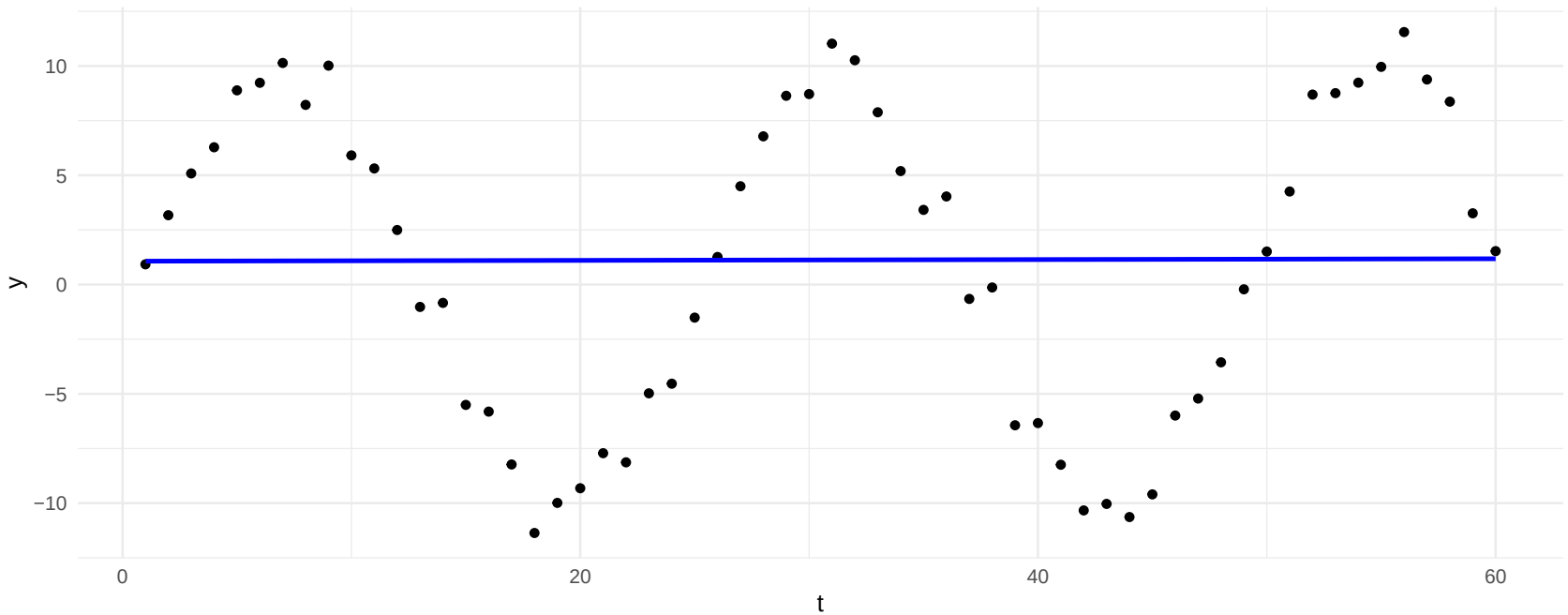


Sine Regression

- Linear regression is a powerful method to understand relationships in data.
- However, it relies on there being some linearity present which is not always true.
- When working with time series, it is likely there will be some patterns that repeat at certain time points.
- Sales of air conditioners are likely to increase every summer, then decrease in the winter, etc.
- Linear regression is not suitable to capture relationships like this.

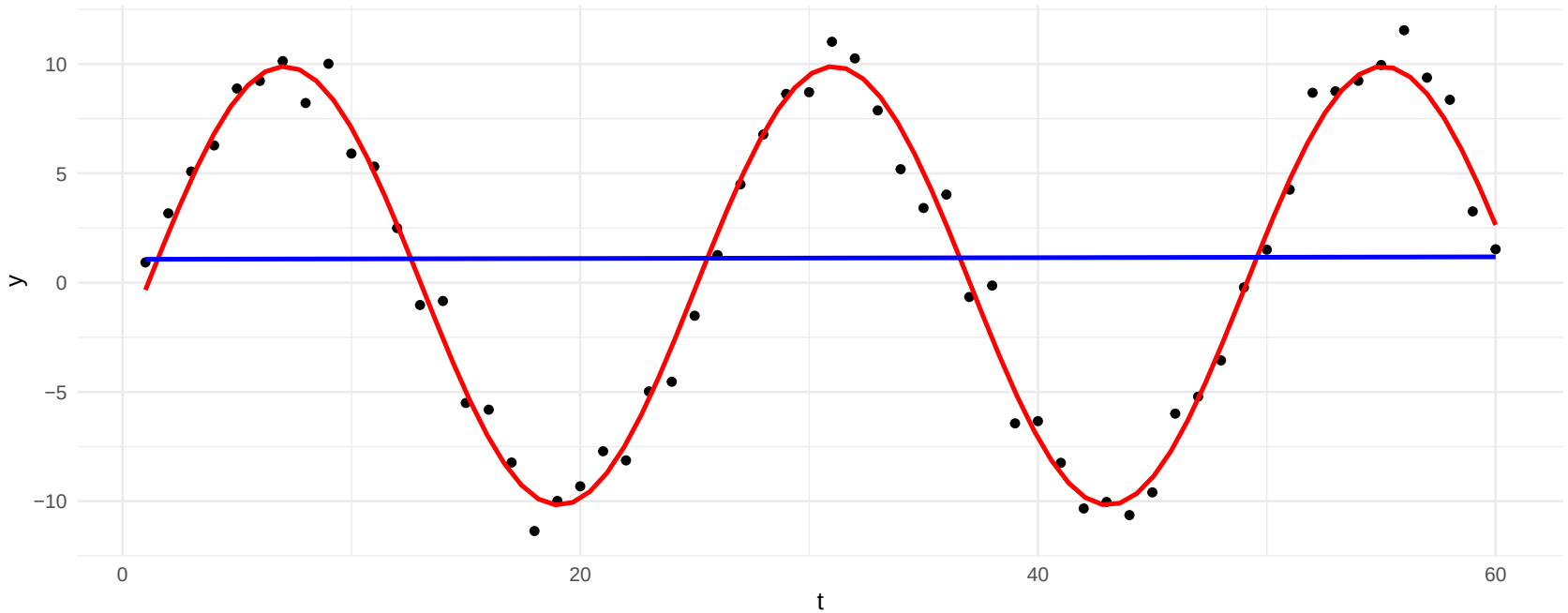
We can easily see this with a quick example. Linear regression is unable to capture the clear structure in the data.

```
dat <- tibble(t = 1:60,
              y = 10 * sin(2 * pi * t / 24 + 6) +
                rnorm(60))
ggplot(dat, aes(t,y)) + geom_point() +
  geom_smooth(method = "lm", color = "blue", se = FALSE)
```



We can get around this by essentially doing some transformation of the data and then fitting a linear regression.

```
fake_plot <- ggplot(dat) + aes(t, y) + geom_point() +
  geom_smooth(formula = y ~ sin(2*pi*x/24) +
                cos(2*pi*x/24), color = "red",
      method = "lm", se=FALSE)
fake_plot + geom_smooth(method = "lm", color = "blue",
                        se = FALSE)
```

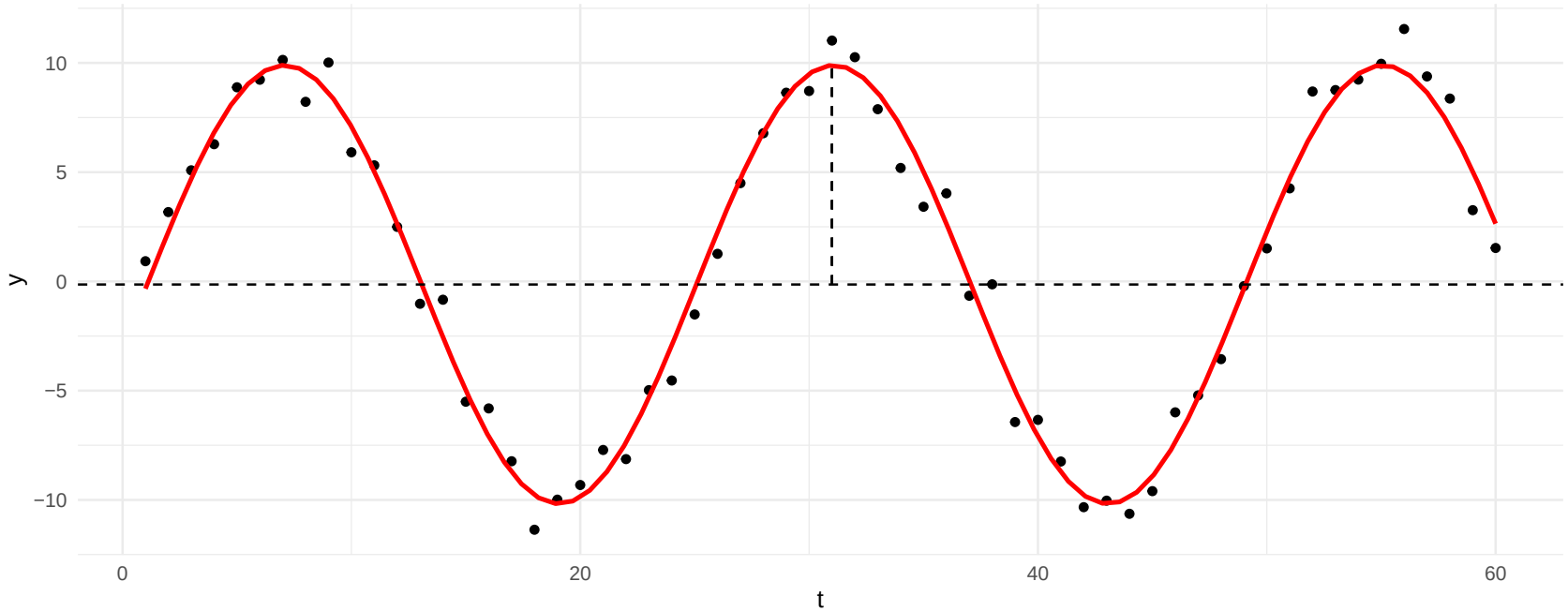


If we use a sinusoid curve to do this, it it parameterized by the amplitude, frequency and the phase.

```
fake_fit <-
  lm(y ~ sin(2*pi*t/24) + cos(2*pi*t/24), dat) %>%
  coef() %>%
tibble(B  = .[1], fi = atan(.[3]/.[2]),
       A  = sqrt(.[2]^2 + .[3]^2))
```

After fitting this model we can look at the fit to the data. Closely recovers the periodic component.

```
fake_plot +
  geom_hline(aes(yintercept = B), data = fake_fit,
             linetype = 2) +
  geom_segment(aes(x = 31, y = B, xend = 31, yend = B +A ),
               data = fake_fit, linetype = 2)
```



Given some knowledge of trigonometry, the above curve looks something like a Sine of Cosine function. One way to do this is to fit the data with a Sine curve

$$Y_t = A \sin(2\pi\omega t + \phi) + B.$$

This has an interpretation, but as it is currently written, it is still not in the form of linear regression.

To do that, we make use of the triginometric identity

$$\sin(\alpha + \beta) = \sin\alpha \cos\beta + \cos\alpha \sin\beta.$$

Using this, then we get

$$Y_t = A\cos\phi\sin(2\pi\omega t) + A\sin\phi\cos(2\pi\omega t) + B.$$

Letting

$$X_1 = \sin(2\pi\omega t), \quad X_2 = \cos(2\pi\omega t),$$
$$\alpha_1 = A\cos\phi, \quad \alpha_2 = A\sin\phi,$$

then we have

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + B.$$

This is a linear regression model in the new variables X_1, X_2.

- Running a linear regression on these new variables will give us estimates $\hat{\alpha}_1, \hat{\alpha}_2, \hat{B}$.
- We want to use these to get estimates of A, ϕ and ω (B is the same in both parameterizations).
- We can again use trigonometry to get back to the original problem.

- We have

$$\alpha_1^2 + \alpha_2^2 = A^2 \left(\cos^2 \phi + \sin^2 \phi \right) = A^2,$$

  so $A = \sqrt{\alpha_1^2 + \alpha_2^2}$.
- Similarly,

$$\frac{\alpha_2}{\alpha_1} = \frac{\sin \phi}{\cos \phi} = \tan \phi,$$

  giving $\phi = \tan^{-1} \left(\frac{\alpha_2}{\alpha_1} \right)$.
- So given $\hat{\alpha}_1, \hat{\alpha}_2, \hat{B}$ we can work back to get $\hat{A}, \hat{B}$ and $\hat{\phi}$.

What do these parameters mean?

- A is the amplitude of the wave.
- B is the overall average.
- ϕ is the phase, an offset term.
- ω is the frequency.

An incomplete list of references.

1. DeGroot, Morris H. Optimal statistical decisions. Vol. 82. John Wiley & Sons, 2005.
2. Freedman, David A. Statistical models: theory and practice. Cambridge University Press, 2009.