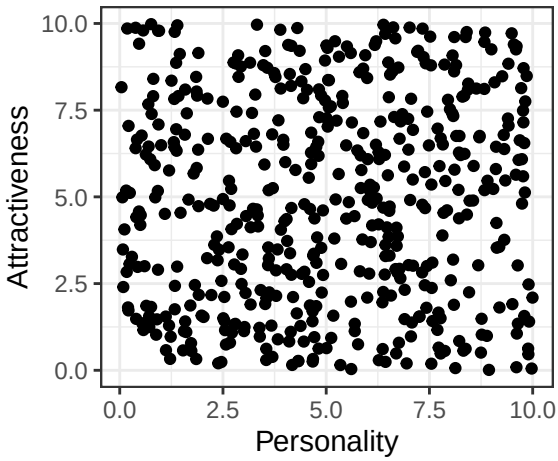
Announcements

Make sure to sign in on the google form.

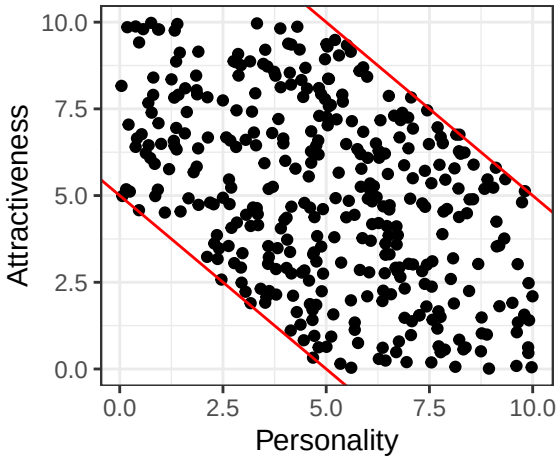
Pset 6...

Prelude: Attractiveness, personality, spurious correlations, and their extensions

With this week's focus on regression, it seemed like a reasonable time to mention an idea I heard a few years ago called Berkson's paradox. Some people say that there's a negative correlation between attractiveness and personality when looking for a romantic partner: people who are attractive can afford to be jerks, but people who aren't attractive need to be nice. However, there could be another explanation. Suppose attractiveness and personality were uncorrelated and each uniform on some 0 to 10 scale. A plot of attractiveness and personality would look like this:



However, you probably wouldn't be interested in someone if they were both unattractive and a jerk. And the people who are both very attractive and very nice are probably already taken. Chop off these two corners of the graph, and voila! Personality and attractiveness are negatively correlated!



The same trend holds (though to a slightly lesser extent) even if only one of the thresholds exists. Such a phenomenon can arise whenever there's a threshold that can be cleared by either of two methods and you're analyzing the correlation among the two methods in only the surviving population (research ability versus teaching ability among faculty, hard work versus intelligence among students admitted to a college, personality versus skill among people with a particular job). Importantly, the correlation implies no causation

at all, and the correlation doesn't even hold when considering the full population; it's purely a result of the threshold. With that, on to some math....

Sine regression

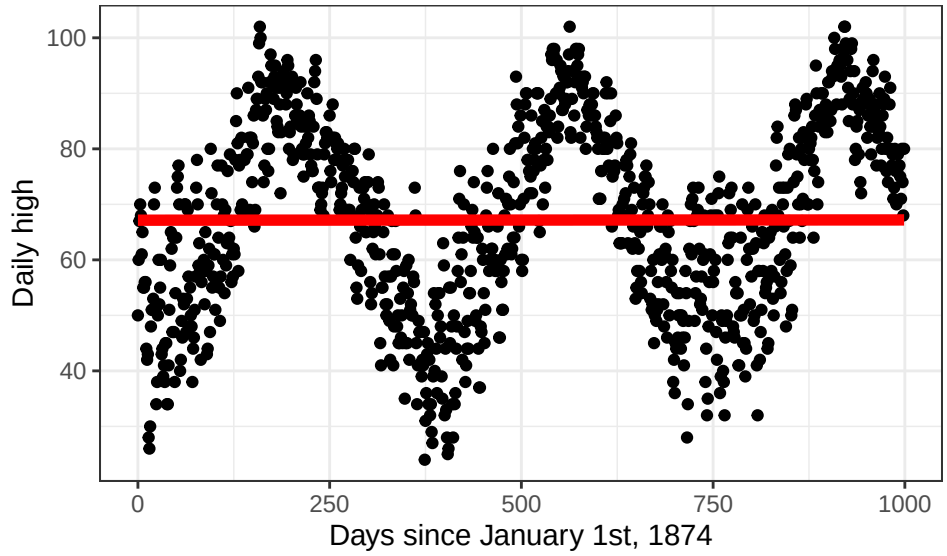
It might seem like it should be impossible to fit sine functions with linear models: a sine equation isn't linear in the parameters or the data. However, a few strategic manipulations can allow linear analysis of sine functions. The following questions deal with data on the daily temperatures from Norfolk, VA available here. Let X_i represent the number of days since January 1st, 1874 (the first day in the dataset) and Y_i represent the maximum temperature on day i.

1. Suppose (extremely) naively that $Y_i = \theta_0 + \theta_1 X_i + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Is this model heteroskedastic or homoskedastic?

2. Provide numerical estimates of θ_0 and θ_1. How well has the model fit?

```
# Fit the model
naive_fit <- lm(tmax ~ day_num, temps)
summary(naive_fit)$coefficients
```

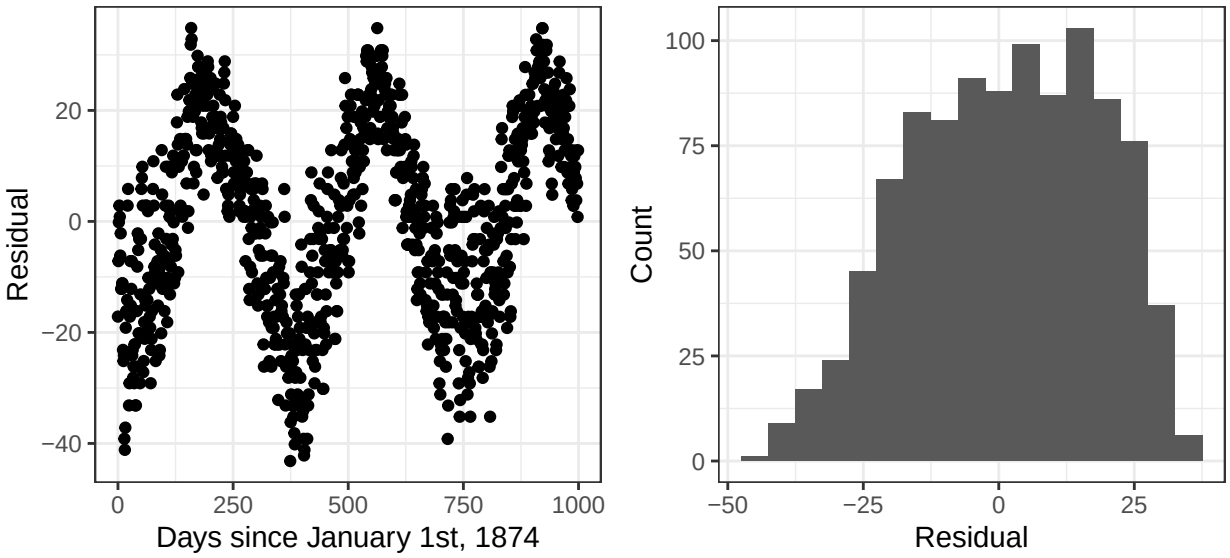
```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 6.714855e+01 1.417947e-01 473.561580 0.000000e+00
## day_num     4.518631e-05 4.558414e-06   9.912726 3.836308e-23
```



3. Provide an approximate 95% confidence interval for how much Norfolk warms per decade on average.

4. Suppose someone used this interval to argue that Norfolk was experiencing climate change. Why should you be skeptical?

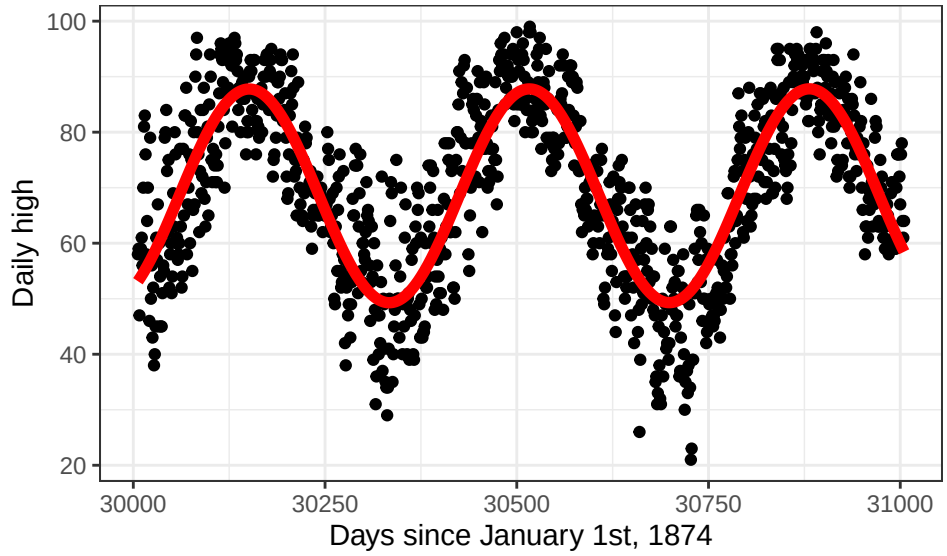
5. Consider the plot of the residuals U_i versus X_i for the first 1000 days. What are the four linear regression assumptions? Which are violated?



6. Consider the model $E(Y_i|X_i) = \beta_0 + \alpha \sin(2\pi\omega(X_i - \phi)) + \beta X_i$. Describe what this model is saying (i.e., what each parameter means). If α, ω, ϕ, and β are unknown, is this a predictive regression? A linear regression? Which of these variables isn't actually unknown?

7. If we take the period fixed as 365.249 days, rewrite the model so it is clearly linear. Find how to determine ϕ and α from your model. The sine addition identity will be useful: $\sin(a+b) = \sin(a)\cos(b) + \sin(b)\cos(a)$.

8. We can fit this linear model and determine $\hat{\alpha}$, $\hat{\phi}$, and $\hat{\beta}$. Interpret the parameters:

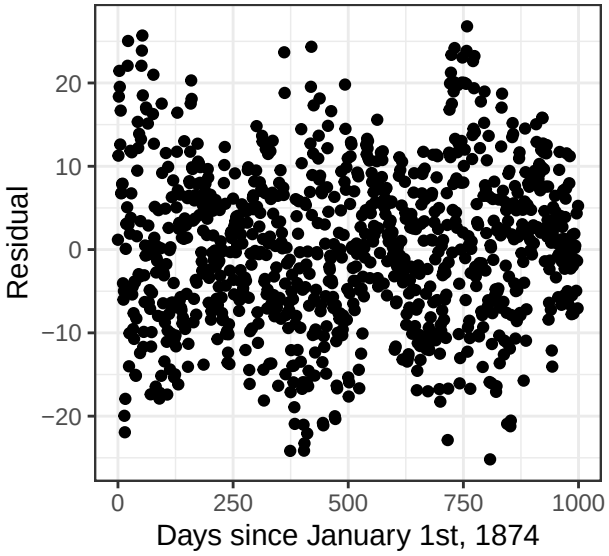


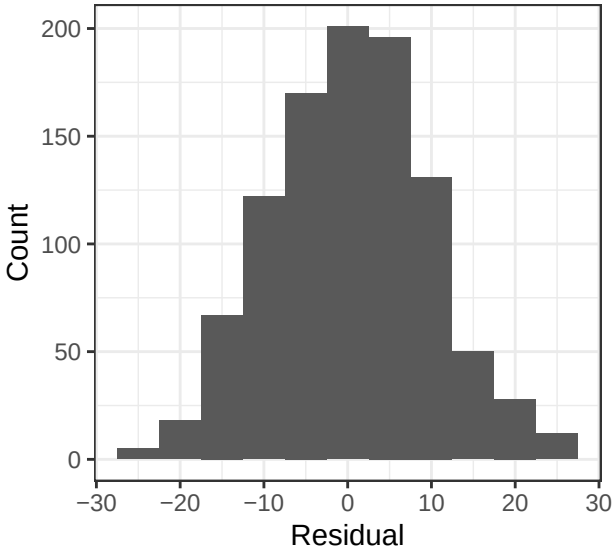

```
##       alpha         phi        beta
##  19.2513717 109.2941907   0.0000452
```

9. Provide a new approximate 95% confidence interval for how much Norfolk warms per decade on average. How does the rate of warming compare to the National Oceanic and Atmospheric Administration's global estimate of 0.14 degrees F per decade since 1880?

```
##                                    Estimate   Std. Error    t value
## (Intercept)                    6.716054e+01 7.964888e-02  843.20755
## day_num                        4.518602e-05 2.560548e-06   17.64701
## sin(2 * pi/365.249 * day_num) -5.860582e+00 5.632012e-02 -104.05840
## cos(2 * pi/365.249 * day_num) -1.833764e+01 5.631560e-02 -325.62265
##                                   Pr(>|t|)
## (Intercept)                   0.000000e+00
## day_num                       1.684586e-69
## sin(2 * pi/365.249 * day_num) 0.000000e+00
## cos(2 * pi/365.249 * day_num) 0.000000e+00
```

10. Using the plot of the residuals U_i versus X_i for the first 1000 days, which assumptions are violated now?





11. Using the regression above, consider the 95% confidence interval for the conditional mean temperature on March 19th 2023 and the 95% prediction interval. How do they compare? The true high was 46. Is it surprising that one interval captured this and the other didn't?

```
predict(lm_fit,
        data.frame("day_num" = as.numeric(as.Date("2023-03-19") - as.Date("1874-01-01"))),
        interval = "confidence", level = 0.95)
```

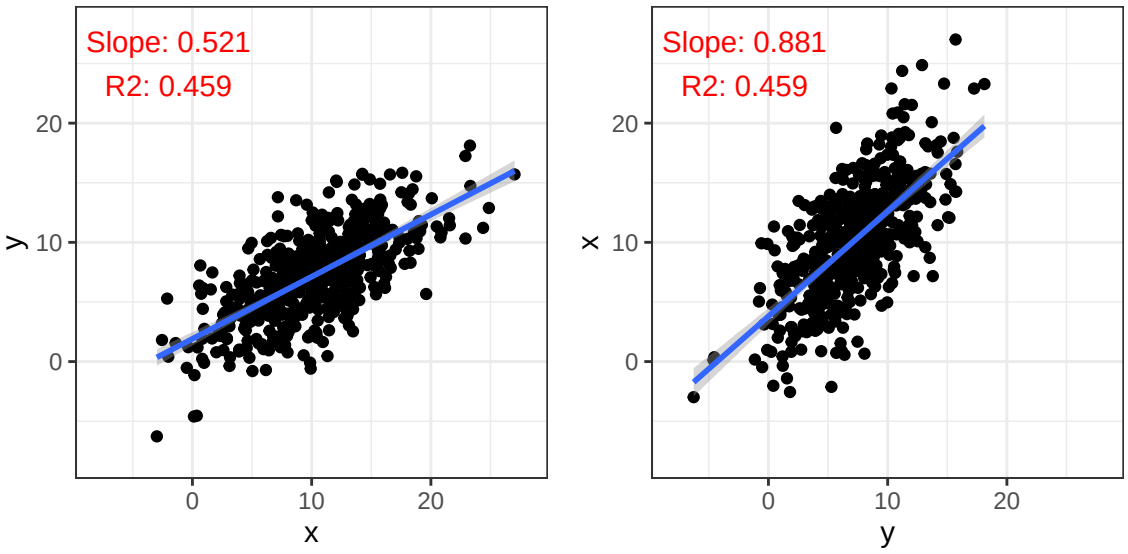
```
##        fit      lwr     upr
## 1 59.16193 58.96865 59.3552
```

```
predict(lm_fit,
        data.frame("day_num" = as.numeric(as.Date("2023-03-19") - as.Date("1874-01-01"))),
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 59.16193 41.04493 77.27893
```

Rule of thumb

Suppose we have n pairs of (X_i, Y_i) and we regress Y on X to get a slope $\hat{\beta}_1$ and X on Y to get a slope $\hat{\beta}_1'$. At first glance, it might seem like the $\hat{\beta}_1 = 1/\hat{\beta}_1'$. However, as you can see in the plots below, this is wrong.



1. Why is this wrong?

2. In the rest of the problem, we'll try to find the proper relationship between the two slopes. Recall that when regressing Y on X, we have

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

Consider our simple regression with the estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and consider the flipped regression estimators

$$\hat{\beta}_1' = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}, \quad \hat{\beta}_0' = \bar{X} - \hat{\beta}_1' \bar{Y}$$

Find an expression for $\hat{\beta}_1'$ in terms of $\hat{\beta}_1$.

3. Solve for R^2 in terms of $\hat{\beta}_1$ and $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. You may use the fact that

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

(See my Stat 111 section 6 notes from last year for why this is the case in simple linear regression.)

4. Use this to write an expression for $\hat{\beta}_1'$ in terms of R^2 and $\hat{\beta}_1$.

Data transformations

In most regressions, right skewed variables are best transformed with a log transformation because this naturally leads to the interpretation that some constant change in the predictor results in a multiplicative change in the output. However, for moderately skewed predictors, a square root transformation can be useful to obtain a better linear model fit. Consider the following two models:

$$Y_i = \beta_0' + \beta_1' X_i + \epsilon_i \tag{1}$$

$$\sqrt{Y_i} = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

for $i \in \{1, ..., n\}$ with $\epsilon_i = \mathcal{N}(0, \sigma^2)$

With $X_i \sim \text{Unif}(0, 10)$, $n = 20$, $\beta_0 = 5$, $\beta_1 = 2$, and $\sigma^2 = 10$, assuming the second model is correct, the following simulation finds the estimates $\hat{\beta}_1$ and $\hat{\beta}_1'$. It also estimates the following quantities: (1) the coverage probability of the 95% confidence interval for $\mu(15)$ based on $\hat{\beta}_1$, (2) the coverage probability of the 95% confidence interval for $\mu(15)$ based on $\hat{\beta}_1'$, (3) the coverage probability of the 95% prediction interval for $Y_{n+1}|X_{n+1} = 5$ based on $\hat{\beta}_1$, and (4) the coverage probability of the 95% prediction interval for $Y_{n+1}|X_{n+1} = 5$ based on $\hat{\beta}_1'$. Interpret the results.

```
set.seed(111)

# Parameters
n <- 20
beta_0 <- 5
beta_1 <- 2
sigma_sq <- 10
nsims <- 10^4

# Vectors for results
mu_covered <- vector(length = nsims)
mu_covered_prime <- vector(length = nsims)
new_covered <- vector(length = nsims)
new_covered_prime <- vector(length = nsims)

for (i in 1:nsims) {
  # Generate data from the model
  x <- runif(n, 0, 10)
  y <- (beta_0 + beta_1 * x + rnorm(n, 0, sqrt(sigma_sq)))^2

  # Fit the model on the original scale
  org_fit <- lm(y ~ x)

  # Fit the true model
  sqrt_fit <- lm(sqrt(y) ~ x)

  # Get the true conditional mean
  mu_true <- (beta_0 + beta_1 * 5)^2 + sigma_sq

  # Create an interval from beta_1 and check coverage
  mu_covered_int <- predict(sqrt_fit, data.frame(x=5),
                            interval = "confidence",
                            level = 0.95)^2
  mu_covered[i] <- mu_true > mu_covered_int[2] &
    mu_true < mu_covered_int[3]
```

```r
  # Create an interval from beta_1 and check coverage
  mu_covered_prime_int <- predict(org_fit, data.frame(x=5),
                                  interval = "confidence",
                                  level = 0.95)
  mu_covered_prime[i] <- mu_true > mu_covered_prime_int[2] & mu_true < mu_covered_prime_int[3]

  # Create a new data point from the model
  x_new <- 5
  y_new  <- (beta_0 + beta_1 * x_new + rnorm(1, 0, sqrt(sigma_sq)))^2

  # Create the intervals and check coverage
  new_covered_int <- predict(sqrt_fit, data.frame(x=x_new),
                                  interval = "prediction",
                                  level = 0.95)^2
  new_covered[i] <- y_new > new_covered_int[2] & y_new < new_covered_int[3]

  new_covered_prime_int <- predict(org_fit, data.frame(x=x_new),
                                  interval = "prediction",
                                  level = 0.95)
  new_covered_prime[i] <- y_new > new_covered_prime_int[2] & y_new < new_covered_prime_int[3]

}

df <- rbind(c(mean(mu_covered), mean(new_covered)),
            c(mean(mu_covered_prime), mean(new_covered_prime)))
colnames(df) <- c("Confidence coverage", "Prediction coverage")
rownames(df) <- c("Beta_1", "Beta_1 prime")
knitr::kable(df)
```

	Confidence coverage	Prediction coverage
Beta_1	0.9342	0.9492
Beta_1 prime	0.7832	0.9616