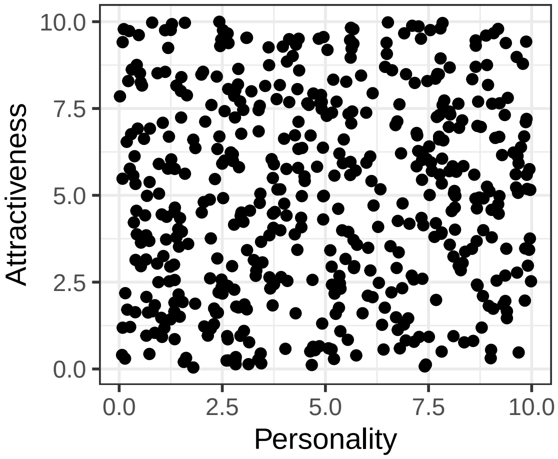
Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

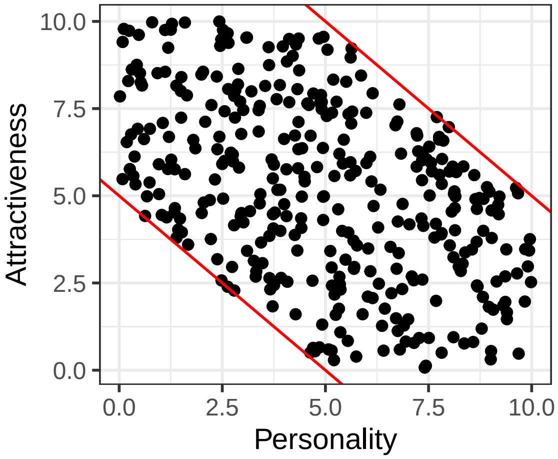
Pset 6 due Friday 3/24

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 (I thought this was from Numberphile but for the life of me I can't track down the source). Some people say that there's a negative correlation between physical 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 (part 3?)

Following Dr. Shephard's sign regression and William Hu's sin regression, this first set of questions will be exploring a case of sine regression that actually has some use. 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?

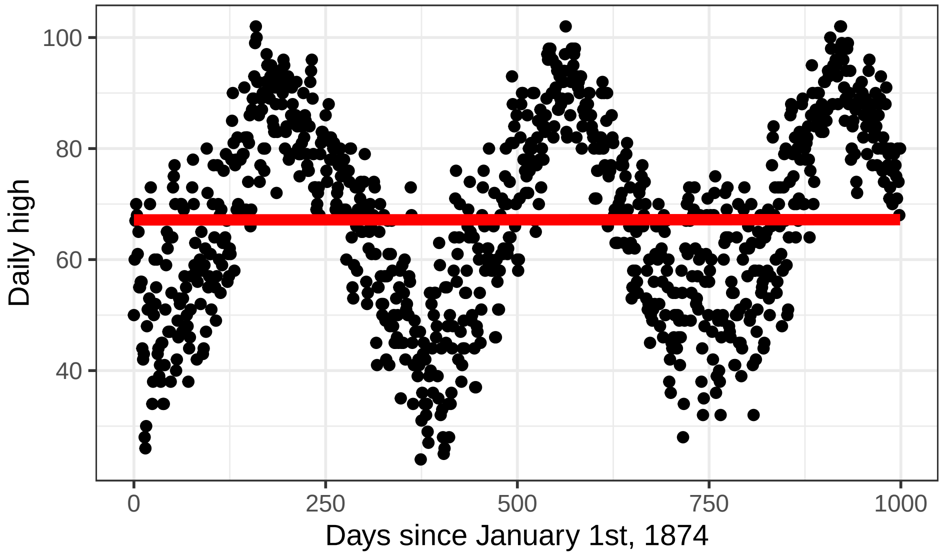
It is homoskedastic because all the conditional variances are the same.

2. Provide numerical estimates of θ_0 and θ_1, the standard error of those estimates, and a plot of the data for the first 1000 days including the fitted line. Hint: the `lm` function takes arguments of the form `y ~ x, df` where `df` is a data frame with columns `y` and `x`. The `summary` command can be used to extract useful information from the fit model.

```
# Read in data and add a days-since-Jan-1-1874 column
temps <- read.csv('data/norfolk_temps.csv')
temps$day_num <- 1:length(temps$tmax) - 1
temps <- temps[!is.na(temps$tmax),]
temps <- temps[,c("tmax", "day_num")]

# Fit the model
naive_fit <- lm(tmax ~ day_num, temps)

# Show fit coefficients
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
```

```
# Get predicted values
predicted_df <- data.frame(tmax = naive_fit$fitted.values, day_num = temps$day_num)

# Plot real versus predicted
ggplot(temps[1:1000,], aes(x = day_num, y = tmax)) +
  geom_point() +
  geom_line(col="red", linewidth=2, data = predicted_df[1:1000,], aes(x=day_num, y=tmax)) +
  theme_bw() +
  xlab("Days since January 1st, 1874") +
  ylab("Daily high")
```



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

Assuming approximate normality, our confidence interval will be the point estimate for $365.249 \cdot 10 \cdot \theta_1$ (the average number of days in a decade times the change per day) plus or minus 1.96 times its standard error:

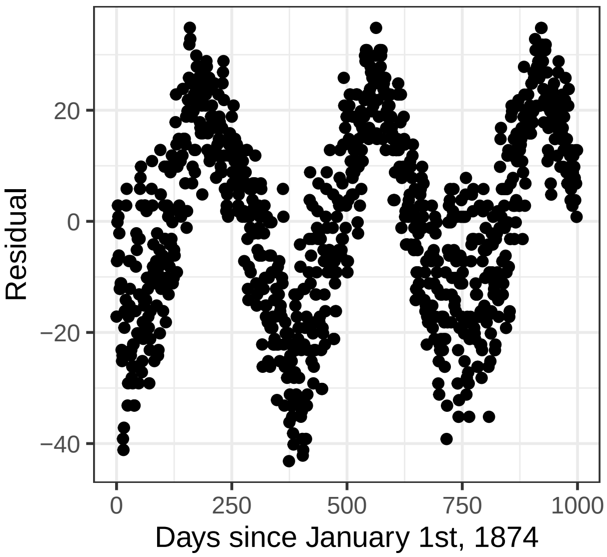
```
10 * 365.249 * c("lb" = summary(naive_fit)$coefficients[2,1] -
                   qnorm(0.975) * summary(naive_fit)$coefficients[2,2],
                 "ub" = summary(naive_fit)$coefficients[2,1] +
                   qnorm(0.975) * summary(naive_fit)$coefficients[2,2])
```

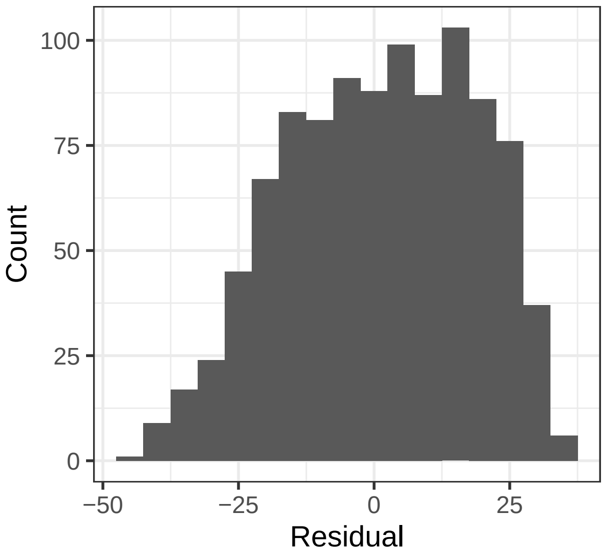
```
##        lb        ub
## 0.1324100 0.1976751
```

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

Though the interval is entirely above 0, the effect is very small, and something as simple as the dataset starting in (cold) January and ending in (hot) July could equally explain this trend.

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?





The four assumptions in linear regression are linearity (the data actually follows a linear model), normality (the error terms are actually Normal), homoskedasticity (the error terms have equal variance), and independence (the Y_i are conditionally independent given $\vec{X}$).

Linearity is definitely not upheld: there is a clear pattern in the data not explained by the linear fit. Normality is also not upheld: the tails are too small for a Normal distribution. Homoskedasticity might be upheld, but there seems to be more variability in the winter than in the rest of the year. Independence is definitely not upheld: temperature data is clearly correlated over time (if it was hot yesterday, it's more likely to be hot today).

6. Consider the model $Y_i = \beta_0 + \alpha \sin(2\pi\omega(X_i - \phi)) + \beta X_i$. 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?

This model is saying that temperatures follow a sin curve with a period $1/\omega$, an offset to the right ϕ, an amplitude α, and an additional change of temperature per day β. This is a predictive regression because we are aiming to find $E(Y_i|X_i = x_i)$. It is not a linear regression because we have unknown ω and ϕ in the sin function, so the model is not linear in the parameters. The parameter ω isn't actually unknown since we know that the period is 365.249 days.

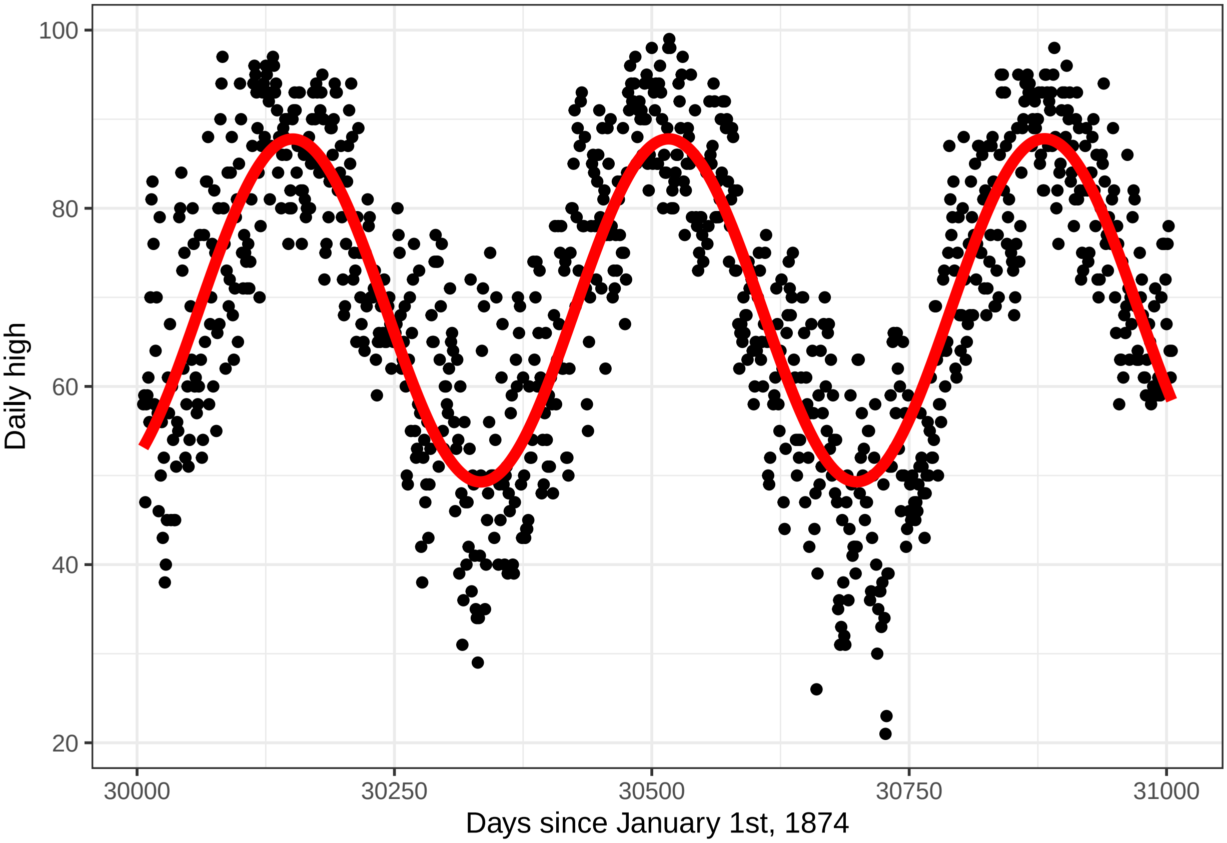
7. If we take the period fixed as 365.249 days, rewrite the model so it is clearly linear. Find how to determine ϕ and α from your model. The sine addition identity will be useful: $\sin(a + b) = \sin(a)\cos(b) + \sin(b)\cos(a)$.

$$Y_i = \beta_0 + \alpha \sin(2\pi\omega(X_i - \phi)) + \beta X_i = \beta_0 + \alpha \cos(-2\pi\omega\phi)\sin(2\pi\omega X_i) + \alpha \sin(-2\pi\omega\phi)\cos(2\pi\omega X_i) + \beta X_i$$

Thus, we will have a linear regression on the known terms $\sin(2\pi\omega X_i)$, $\cos(2\pi\omega X_i)$, and X_i, and the coefficients will be $\beta_0 = \beta_0$, $\beta_1 = \beta$, $\beta_2 = \alpha\cos(-2\pi\omega\phi)$, $\beta_3 = \alpha\sin(-2\pi\omega\phi)$. We can find α with $\alpha = \sqrt{\beta_2^2 + \beta_3^2}$ and $\phi = \cos^{-1}(\beta_2/\alpha)/(2\pi\omega)$.

8. Fit this linear model and determine $\hat{\alpha}$, $\hat{\phi}$, and $\hat{\beta}$.

```r
# Fit the model
lm_fit <- lm(tmax ~ day_num + sin(2 * pi / 365.249 * day_num) +
               cos(2 * pi / 365.249 * day_num), temps)

# Get the predicted values from the model
predicted_df <- data.frame(tmax = lm_fit$fitted.values, day_num = temps$day_num)

# Plot the actual and predicted values
ggplot(temps[30001:31000,], aes(x=day_num, y=tmax)) +
  geom_point() +
  theme_bw() +
  geom_line(color = "red", linewidth = 2, data = predicted_df[30001:31000,],
            aes(x=day_num, y=tmax)) +
  xlab("Days since January 1st, 1874") +
  ylab("Daily high")
```



```
# Calculate parameters
alpha <- sqrt(lm_fit$coefficients[3]^2 + lm_fit$coefficients[4]^2)
phi = acos(lm_fit$coefficients[3]/alpha) *
  365.249 / (2 * pi)
params <- c(alpha, phi, lm_fit$coefficients[2])
names(params) <- c("alpha", "phi", "beta")

# Print parameter estimates
params
```

```
##        alpha          phi         beta
## 1.925137e+01 1.092942e+02 4.518602e-05
```

April 19th (day 109) is the offset, which corresponds to the spring day with the most average temperature of the year. The amplitude is 19.2, so the average mid-winter high is about 38.4 degrees colder than the average mid-summer high The coefficient $\beta = 4.52 \cdot 10^{-5}$ shows that the temperature warms by $4.52 \cdot 10^{-5}$ degrees per day on average.

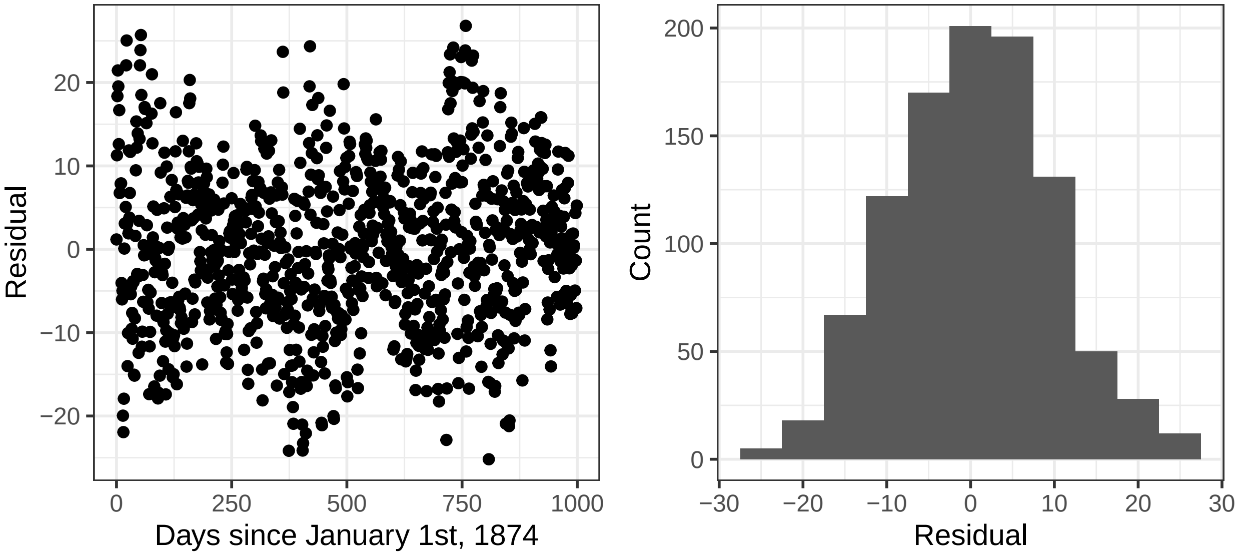
9. Provide a new approximate 95% confidence interval for how much Norfolk warms per decade on average. How does the rate of warming compare to the National Oceanic and Atmospheric Administration's global estimate of 0.14 degrees F per decade since 1880?

```
10 * 365.249 * c("lb" = summary(lm_fit)$coefficients[2,1] - qnorm(0.975) *
                   summary(lm_fit)$coefficients[2,2],
                 "ub" = summary(lm_fit)$coefficients[2,1] + qnorm(0.975) *
                   summary(lm_fit)$coefficients[2,2])
```

```
##        lb        ub
## 0.1467112 0.1833718
```

The confidence interval here slightly misses the NOAA number, but the NOAA number is for ocean and surface temperatures combined over a much larger area, so this is reasonable. Notably, this interval is about half as wide as before and not subject to the same issues with start and end points.

10. Make a plot of the residuals U_i versus X_i for the first 1000 days. Which assumptions are violated now?



Linearity is much better now: there is much less of a pattern in the residuals. Normality is very well upheld now (a QQ plot of this data is actually one of the best I've ever seen). Homoskedasticity still seems problematic with more variability in the winter than in the rest of the year. Independence is still not upheld: even once we account for the day in the year, weather trends mean that days will be correlated (you can see this well if you reduce the number of residuals plotted).

11. Using the regression above, a 95% confidence interval has been calculated for the conditional mean temperature on March 19th 2023. Now, provide a 95% prediction interval. How does this compare to the confidence interval? 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
```

Both have the same mean, but this interval is much wider. As expected, this interval captures the observed temperature, but the previous interval didn't.

Rule of thumb

1. The coefficient of determination R^2 for a model roughly measures how much variance in the outcome is explained by the predictor. This is often reported as a measure of how good a model is, and it can be written mathematically as
$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$
where RSS is the residual sum of squares: $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. Consider the model
$$Y_i = \theta_{0,Y\sim X} + \theta_{1,Y\sim X} X_i + \epsilon_i$$
with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Suppose we only have our usual OLS estimators
$$\hat{\theta}_{1,Y\sim X} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}, \quad \hat{\theta}_{0,Y\sim X} = \bar{Y} - \hat{\theta}_{1,Y\sim X}\bar{X}$$
and R^2 but we actually want to estimate the opposite effect: $\hat{\theta}_{1,X\sim Y} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$. (This scenario is quite plausible since papers with a linear model will usually report the best fit slope as well as the model's R^2 even if they don't publish all the raw data.) Explain intuitively why $\hat{\theta}_{1,Y\sim X} \neq 1/\hat{\theta}_{1,X\sim Y}$.

As we'll see in this week's homework and in the next part of this question, regression to the mean ensures that the product of the effects will be shrunk slightly towards 0 away from 1.

2. Find an expression for $\hat{\theta}_{1,X\sim Y}$ in terms of $\hat{\theta}_{1,Y\sim X}$ assuming you had access to all the data.
$$\hat{\theta}_{1,X\sim Y} = \hat{\theta}_{1,Y\sim X} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

3. Show that $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ starting with the fact that $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$.

Squaring each side and summing, we get:
$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2\sum_{i=1}^n (Y_i - \hat{Y})(\hat{Y}_i - \bar{Y})$$
so it remains to show that $\sum_{i=1}^n (Y_i - \hat{Y})(\hat{Y}_i - \bar{Y}) = 0$. Using $\hat{\theta}_1 = \hat{\theta}_{1,Y\sim X}$ and $\hat{\theta}_0 = \hat{\theta}_{0,Y\sim X}$,
$$\begin{aligned}
\sum_{i=1}^n (Y_i - \hat{Y})(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n (Y_i - (\hat{\theta}_0 + \hat{\theta}_1 X_i))((\hat{\theta}_0 + \hat{\theta}_1 X_i) - \hat{\theta}_0 + \hat{\theta}_1 \bar{X}) \\
&= \hat{\theta}_1 \sum_{i=1}^n (Y_i - (\hat{\theta}_0 + \hat{\theta}_1 X_i))(X_i - \bar{X}) \\
&= \hat{\theta}_1 \sum_{i=1}^n (Y_i - (\bar{Y} - \hat{\theta}_1 \bar{X} + \hat{\theta}_1 X_i))(X_i - \bar{X}) \\
&= \hat{\theta}_1 \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) + \hat{\theta}_1 (\bar{X} - X_i)(X_i - \bar{X}) \\
&= \hat{\theta}_1 \left(\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) \right) - \hat{\theta}_1 \left(\hat{\theta}_1 \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\
&= 0
\end{aligned}$$

3. Solve for R^2 in terms of $\hat{\theta}_{1,Y\sim X}$ and $\frac{\sum_{i=1}^n (X_i-\bar{X})^2}{\sum_{i=1}^n (Y_i-\bar{Y})^2}$. The fact that $\bar{Y} = \hat{\theta}_{0,Y\sim X} + \bar{X}\hat{\theta}_{1,Y\sim X}$ may be useful (from rearranging the equation for $\hat{\theta}_{0,Y\sim X}$).

With $\hat{\theta}_1 = \hat{\theta}_{1,Y\sim X}$ and $\hat{\theta}_0 = \hat{\theta}_{0,Y\sim X}$ as before:

$$\begin{aligned}
R^2 &= 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\bar{Y} - (\hat{\theta}_0 + \hat{\theta}_1 X_i))^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{\theta}_0 + \hat{\theta}_1 \bar{X} - (\hat{\theta}_0 + \hat{\theta}_1 X_i))^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \frac{\sum_{i=1}^n (\hat{\theta}_1 \bar{X} - \hat{\theta}_1 X_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\
&= \hat{\theta_1}^2 \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}
\end{aligned}$$

4. Use this to write an expression for $\hat{\theta}_{1,X\sim Y}$ in terms of R^2 and $\hat{\theta}_{1,Y\sim X}$.

$$\hat{\theta}_{1,X\sim Y} = \frac{R^2}{\hat{\theta}_{1,Y\sim X}}$$

Notably, $0 \leq R^2 \leq 1$ for an OLS model, so $\hat{\theta}_{1,X\sim Y} \cdot \hat{\theta}_{1,Y\sim X} \leq 1$, which gives another view of regression to the mean.

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.

```r
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]
  # 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

Using the correct model and transforming the intervals to the original scale gives the nominal coverage probability for the prediction interval but not for the confidence interval. Fitting the wrong model on the original scale results in both being off.