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

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?

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")]

# TODO: Fit the model

# TODO: Show fit coefficients

# TODO: Get predicted values

# TODO: Plot real versus predicted
```

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

```
# TODO: Find interval
```

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 (switch `eval=T` in the next chunk once you've fit `naive_fit`). What are the four linear regression assumptions? Which are violated?

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?

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

```
# TODO: Fit the model
lm_fit <- lm()

# TODO: Calculate parameters
```

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?

```
# TODO: Find interval
```

10. Make a plot of the residuals U_i versus X_i for the first 1000 days (switch `eval=T` in the next chunk once you've fit `lm_fit`). Which assumptions are violated now?

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)
```

```
# TODO: Calculate prediction interval
```

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 $\hat{\theta}_{1,Y\sim X} \neq 1/\hat{\theta}_{1,X\sim Y}$.

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.

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 $(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$.

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 $\hat{\theta}_{0,Y\sim X}$).

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

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]
```

```
  # 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