Log Transformations

It is pretty common to see a distribution of data that is extremely right skewed. For example, individual level income and county level population measures are both highly right skewed. Both have many observations with low income/population, and a few observations with high income/population.

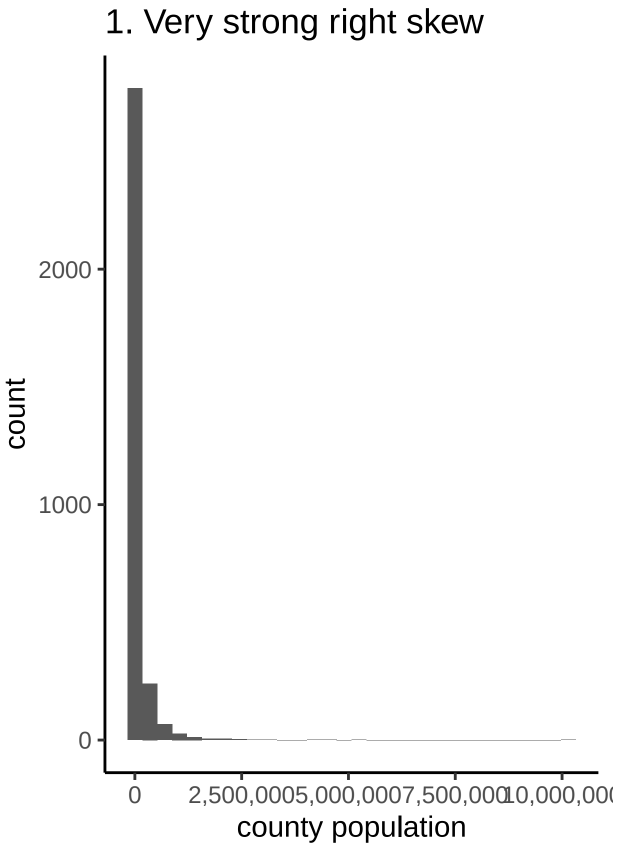
This kind of skewed data results in outliers, or data points that are unusually far away from bulk of the data. As a result

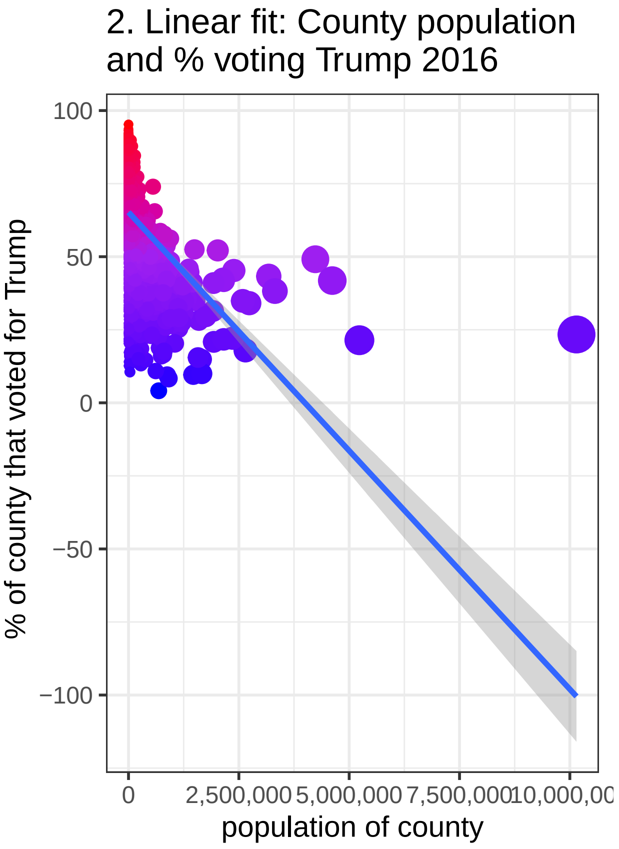
- the regression line is always pulled towards the outliers
- the skew of the data can, but does not always, indicate a nonlinear relationship

We'll see how a log transformation of the highly right skewed variable will fix either or both issues. Note that severely left skewed data is often fixed using a square, cube, or exponential transformation. The goal is to obtain an approximately normally distributed variable for any continuous data; ask your professor if you encounter a continuous variable that you are having trouble transforming into a normal distribution.

The next two graphics show

1. A histogram of a severely right skewed variable, popEstimate16
2. What a scatterplot and linear regression line looks like for population plotted against the vote for Trump in the county data.


1. Very strong right skew


2. Linear fit: County population and % voting Trump 2016

A log transformation reduces the impact of these outliers in two ways

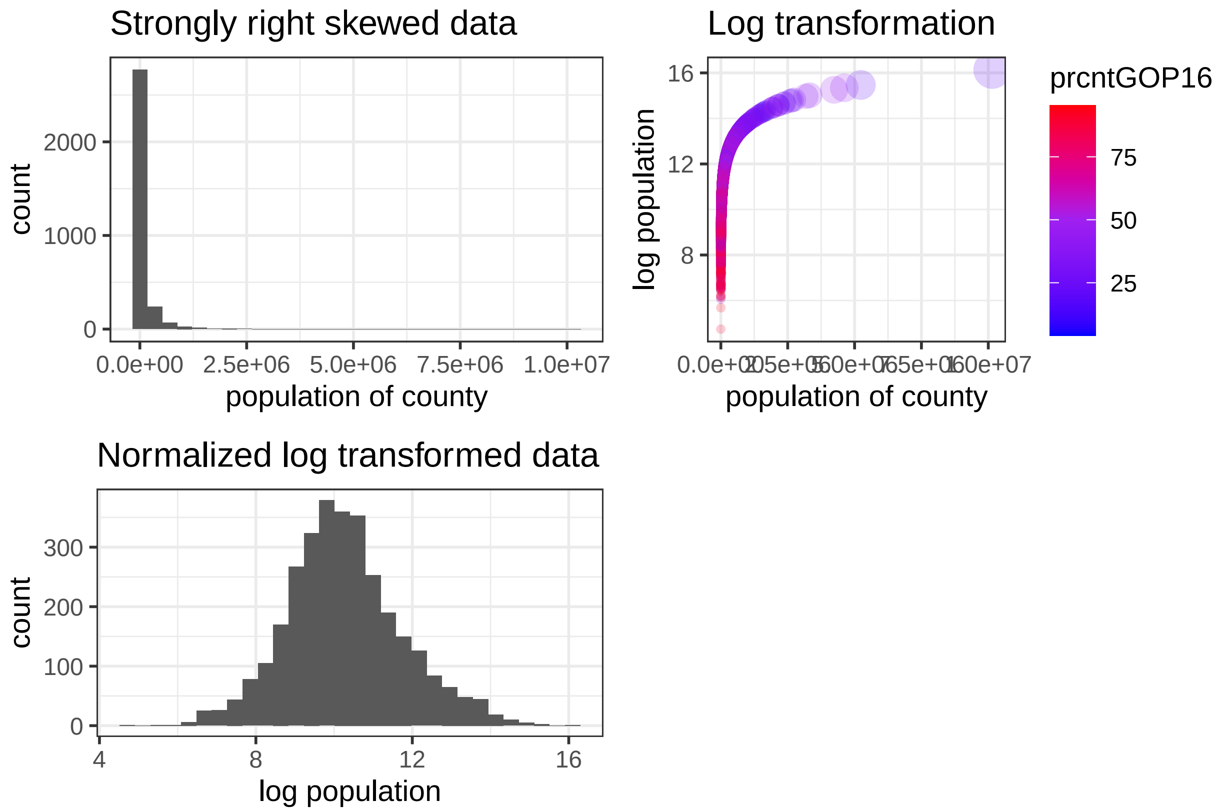
- expanding the distance between the small values and shrinking the distance between the large values
- allowing for a curved relationship.

Creating a log transform is simple, using either of the two methods in the following R chunk.

```
#tidy code, lets you set where the new variable appear
countyData <- countyData %>%
  mutate(log_pop = log(popEstimate16), .after = popEstimate16)

## same operation in base R code, new variable always appears as the last column
# countyData$log_pop <- log(countyData$popEstimate16)
```

You can see that the data goes from right skewed for the population variable to normally distributed for the logged population variable.



County Demographics & Vote

In general, the linear relationships are best plotted using variables that are normally distributed, as much as possible. But before we see the linear relationship, let's look at the log relationship plotted with the original data.

Here, I create the plots with the curved log relationship for the original, untransformed variables, and compare them with the linear relationship with the original data. Note that the best fit line for the log relationship falls very nicely in the center most vertical cross-sections of the data. As described in the lab on Anscombe's quartet, *any* well fitting regression line should approximately fit any given vertical cross-section of a scatterplot. The zoomed in graphic shows this more clearly.

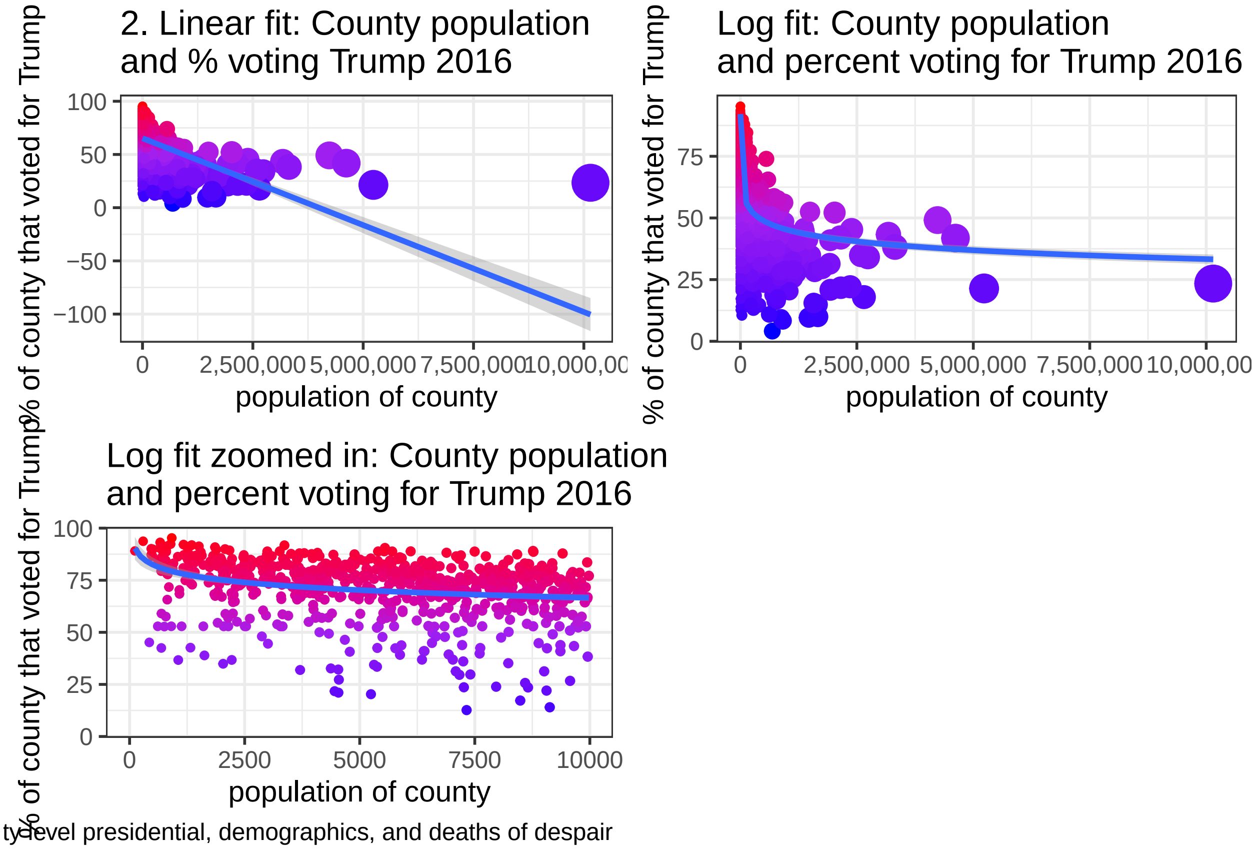
Specifically, the curved relationship being plotted below is represented using the following equation.

```
  y = beta1 * log(x1)  + beta0
```

```
% Trump = beta1 * log(population) + beta0
```

Repeated in fancy math notation:

$$
\begin{aligned}
\text{y} &= \beta_1 * log(x) + \beta_0 \\
\text{Percent Trump} &= \beta_1 * \log(\text{population}) + \beta_0
\end{aligned}
$$



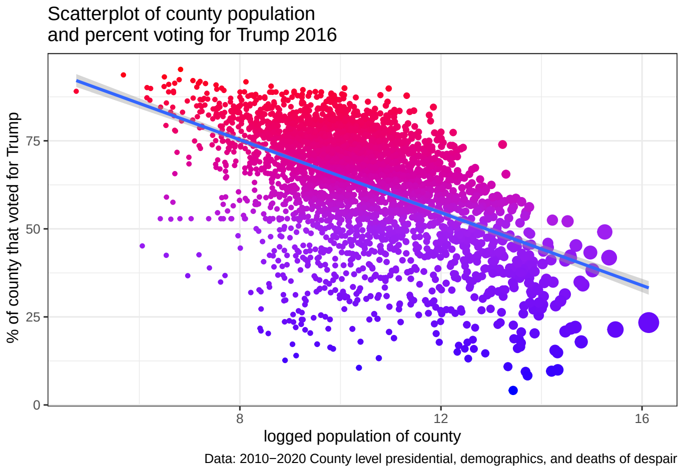
However, for better or worse, we normally do not show the curved log fit line. Instead we generally transform the x variable itself and estimate a *linear* fit. Specifically, we estimate the following relationship.

```
% Trump = beta1 * log_pop  + beta0
```

Repeated in fancy math notation:

$$
\begin{aligned}
\text{y} &= \beta_1 * x + \beta_0 \\
\text{Percent Trump} &= \beta_1 * \text{logged population} + \beta_0
\end{aligned}
$$

This is mathematically identical to plotting the curved log fit, but is easier to compute and pleases mathematicians who find linear equations easier to handle than non-linear equations. The resulting linear fit line, plotted with the *logged* population variable, looks like this.


Scatterplot of county population
and percent voting for Trump 2016

Data: 2010−2020 County level presidential, demographics, and deaths of despair

The main point here is that any time you see a highly skewed variable, you should consider transforming it to be normally distributed. The log transformation generally fixes right skews. Square, cube, or exponential transformations tend to fix left skews. Ask your professor for help with transforms if they seem needed and a simple transform (square or log) doesn't seem to improve the fit of your regression model to the data.

Exercise 3: 1 pt

Create a histogram of faminc_num in CCES.

Answer (code):

Grade: /1

Exercise 4: 2 pts

Plot the relationship between faminc_num and race_attitudes in CCES, where faminc_num is the x variable and race_attitudes is the y variable. You'll need to add the layers geom_jitter and geom_smooth(method = lm) to your ggplot() object. Feel free to color the graphic or not. If you choose to apply a sizing variable, you'll want to use the variable commonweight to size the dots. The commonweight variable puts more weight on respondents who are less likely to respond to the survey. For example, a low education white male is much less likely to respond than a high education white female, and therefore the survey mitigates the overresponse of white females by giving them lower weights. Ask us if you have any questions about weights in your data.

Answer (code):

Grade: /2

The following code creates a new variable called log_faminc_num in CCES.

```r
cces <- cces %>% mutate(log_faminc_num = log(faminc_num), .after = faminc_num)
```

Exercise 5: 1 pt

Create a histogram of log_faminc_num.

Answer (code):

Grade: /1

Exercise 6: 1 pt

Is it more normally distributed than faminc_num?

Answer (a few words or a sentence):

Grade: /1

Exercise 7: 2 pts

Plot the relationship between log_faminc_num and race_attitudes in CCES, where log_faminc_num is the x variable and race_attitudes is the y variable.

Answer (code):

Grade: /2

Exercise 7: 2 pts

Does the linear fit seem better (multiple answers are possible here, use your judgement and explain why)? Can you see the spread between the low income values better?

Answer (a few sentences):

Grade: /2

Exercise 8: 2 pts

What is the overall trend between income and racial attitudes. That is, do high income folk have higher or lower racial attitude scores than low income folk?

Answer (a sentence or two):

Grade: /2