Analysis of gas

2019-04-10

Data Prep

1. Dropped records with missing gas values at t1 and t2. Call the resulting data set `df`.
2. Separated `df` into two subsets:
   - `df_tiny`: gas < 1000 at t1 or t2.
   - `df_main`: gas >= 1000 at t1 and t2
3. Created long-format version:
   - `df_long`: long-format version of the full set `df`
   - `df_tiny_long`: long-format version of the subset `df_tiny`
   - `df_main_long`: long-format version of the subset `df_main`

Analyze the full set `df_tiny`

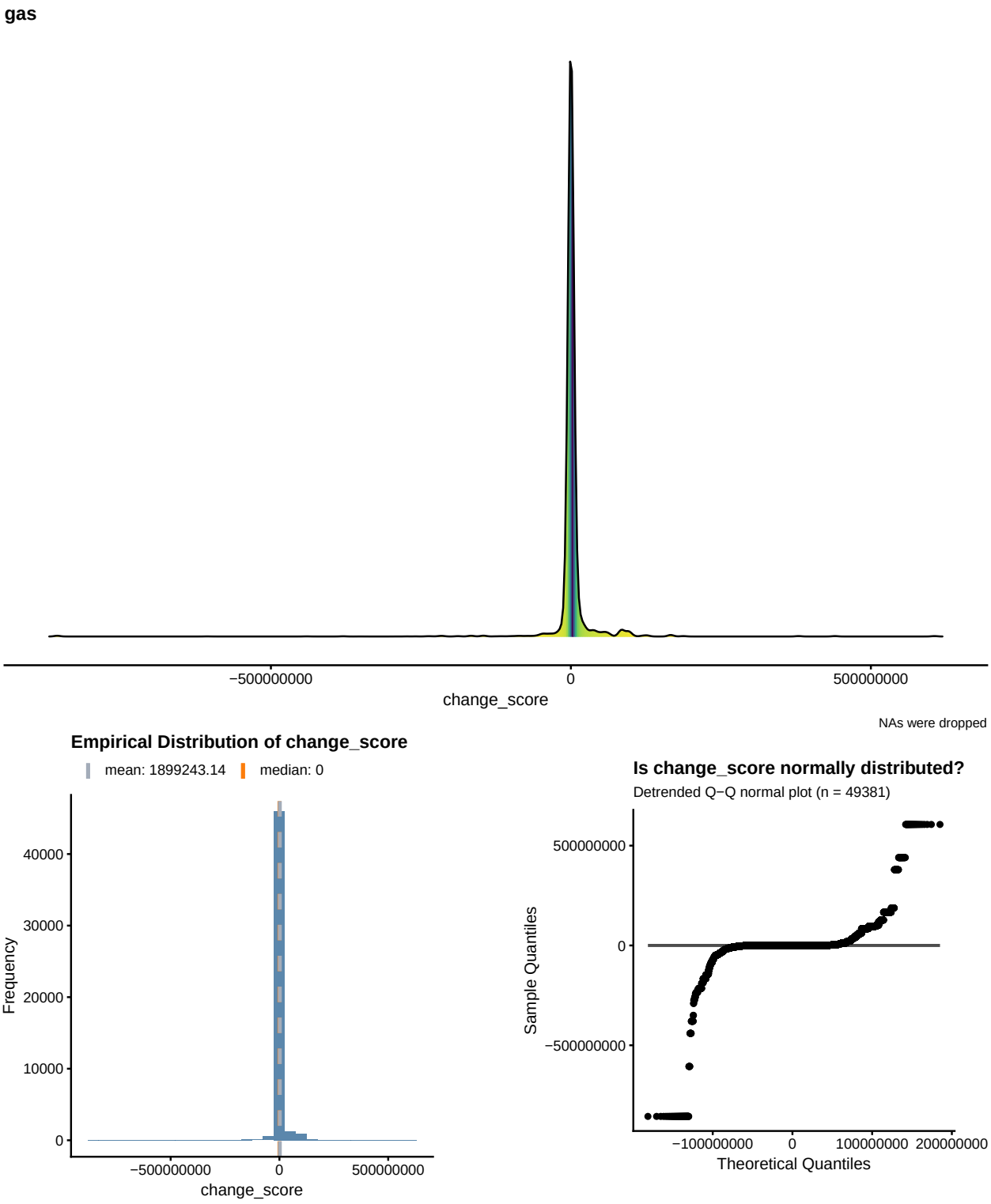
Table 1: Sample Summary Statistics of gas

time	n_tribes	n	mean	median	std	skewness	kurtosis	SEM
t1	179	49381	3227196	0	33474955	20.24567	477.5600	150639.9
t2	179	49381	5126439	0	26344294	12.49777	223.1352	118551.4

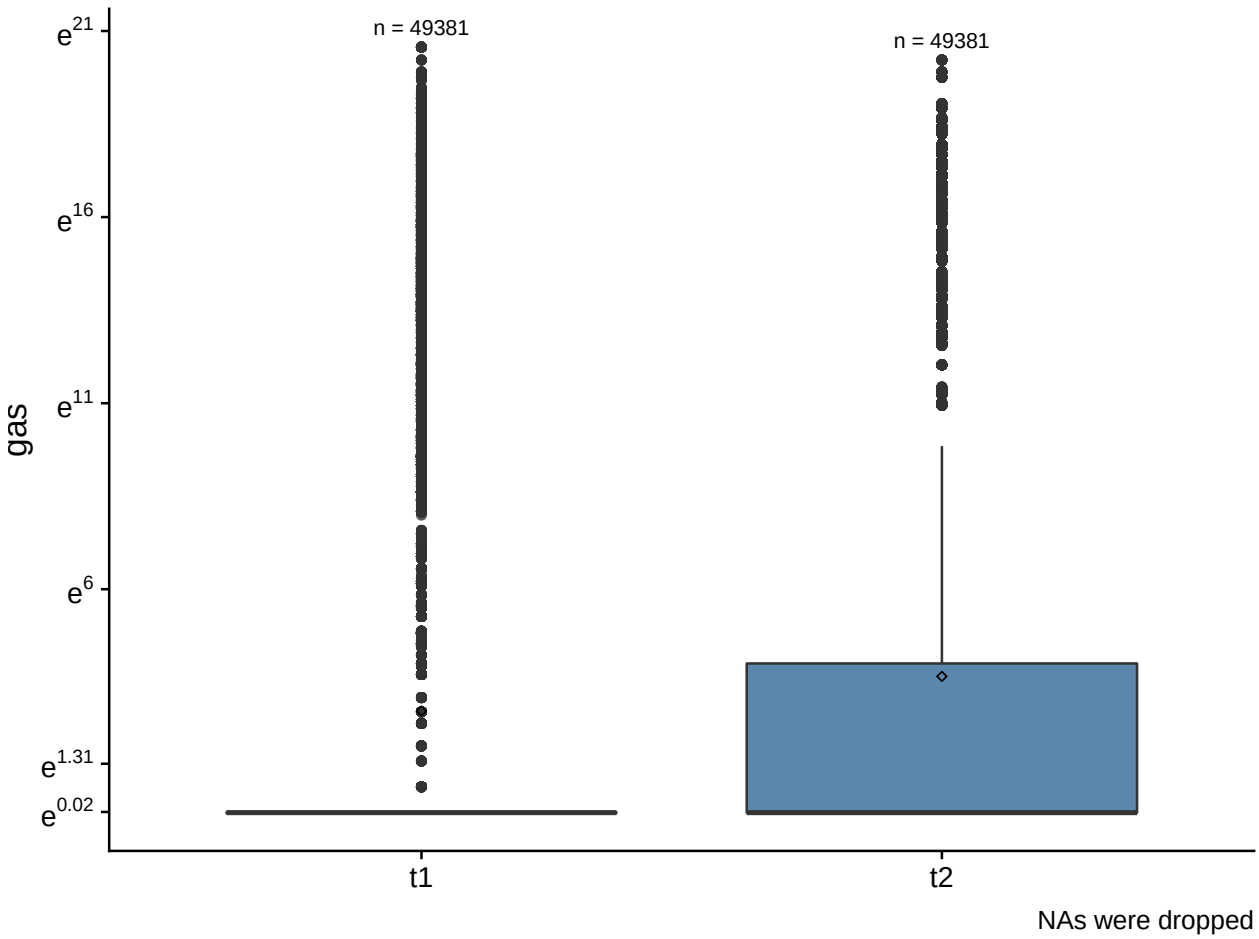
Q1. Is there a difference between t1 and t2?

Descriptive Analysis

First we look at the distribution of the change scores between t2 and t1 (t2 - t1). The following density plot, histogram and detrended qqnormal plot show that its distribution is symmetric but not normal, with very long tails extending far in the positive and negative directions. A normal distribution would produce a detrended qqnormal plot with most of the data points randomly scattered around the line `y=0` and within the grayish blue confidence band. There's a second smaller peak along the right tail, but the left tail is longer than the right tail. The sample median is 0. The sample mean is a large positive number, but is relatively small in relation to the extreme tail values. This can be seen from the histogram where the silver dashed line (mean) almost overlap with the orange dashed line (median = 0). As a result, we cannot tell from these plots if the mean gas change from t1 to t2 is significantly different from zero.



Next we look at the distributions of (log1p transformed) gas values at t1 and t2 side by side with regular boxplots (cannot draw notched boxplot because of the zero values). We see there're either extremely tiny or large values at t1, so much so that there isn't a box body. At t2, there's a box body with lots of large outliers. So we cannot tell if there's an significant difference in the population mean gas values between t2 and t1.



Statistical Analysis

To test if there's a difference between the population mean gas values at t1 and t2, we ran an one-way repeated measure ANOVA and found that the term `time` is NOT significant with a p-value of 0.128. We also ran a linear mixed model with tribe as random effect and found that `time` is highly significant with a p-value < 0.0001. Because the p-value under ANOVA is not too big, and it's borderline significant under the significance level of 0.1, we'll cautiously conclude there's a significant difference between the mean gas values at t2 and t1.

When reading the output from ANOVA and linear mixed model, you want to focus on the reported p-value of the term `time`. It tells you the probability of observing a difference between t1 and t2 as extreme as in the sample data due to chance or randomness. If it's small, it's more likely that the observed difference is not due to chance. To decide how small is "small", the convention is to compare the p-value with 0.05. (But you don't have to use 0.05, it's really your choice. For example, 0.01 or 0.1 are also commonly used in different applications). If it's less than 0.05, we say the observed difference between t1 and t2 in the sample is likely not due to chance and hence can be generalized to the entire population. In other words, the difference is (statistically) significant. Otherwise when the p-value is greater than 0.05, we say the observed difference between t1 and t2 in the sample is likely a fluke and cannot be generalized to the entire population. In other words, the difference is not (statistically) significant.

```
1-way Repeated Measure ANOVA Output:

Error: tribe
          Df              Sum Sq           Mean Sq F value Pr(>F)
```

```
Residuals 178 7110549670751643648 39946908262649680

Error: tribe:time
           Df               Sum Sq           Mean Sq F value Pr(>F)
time        1   89061707170019088 89061707170019088   2.341  0.128
Residuals 178 6772764354417592320 38049237946166248

Error: Within
             Df                Sum Sq         Mean Sq F value Pr(>F)
Residuals 98404 75721359457288667136 769494730471207
```

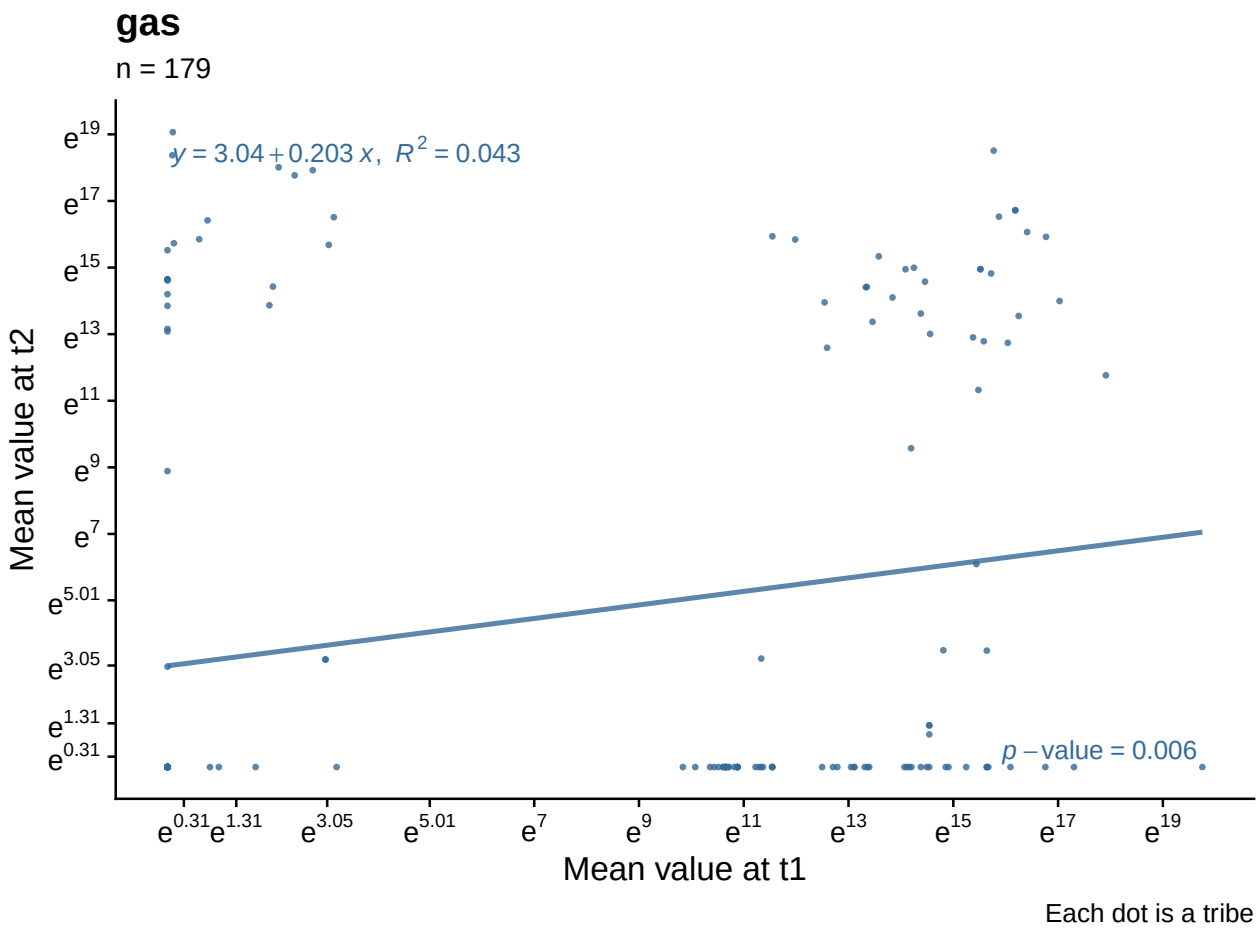
```
Linear Mixed Model Output:
            numDF denDF  F-value p-value
(Intercept)     1 98582  17.7812  <.0001
time            1 98582 106.4126  <.0001
```

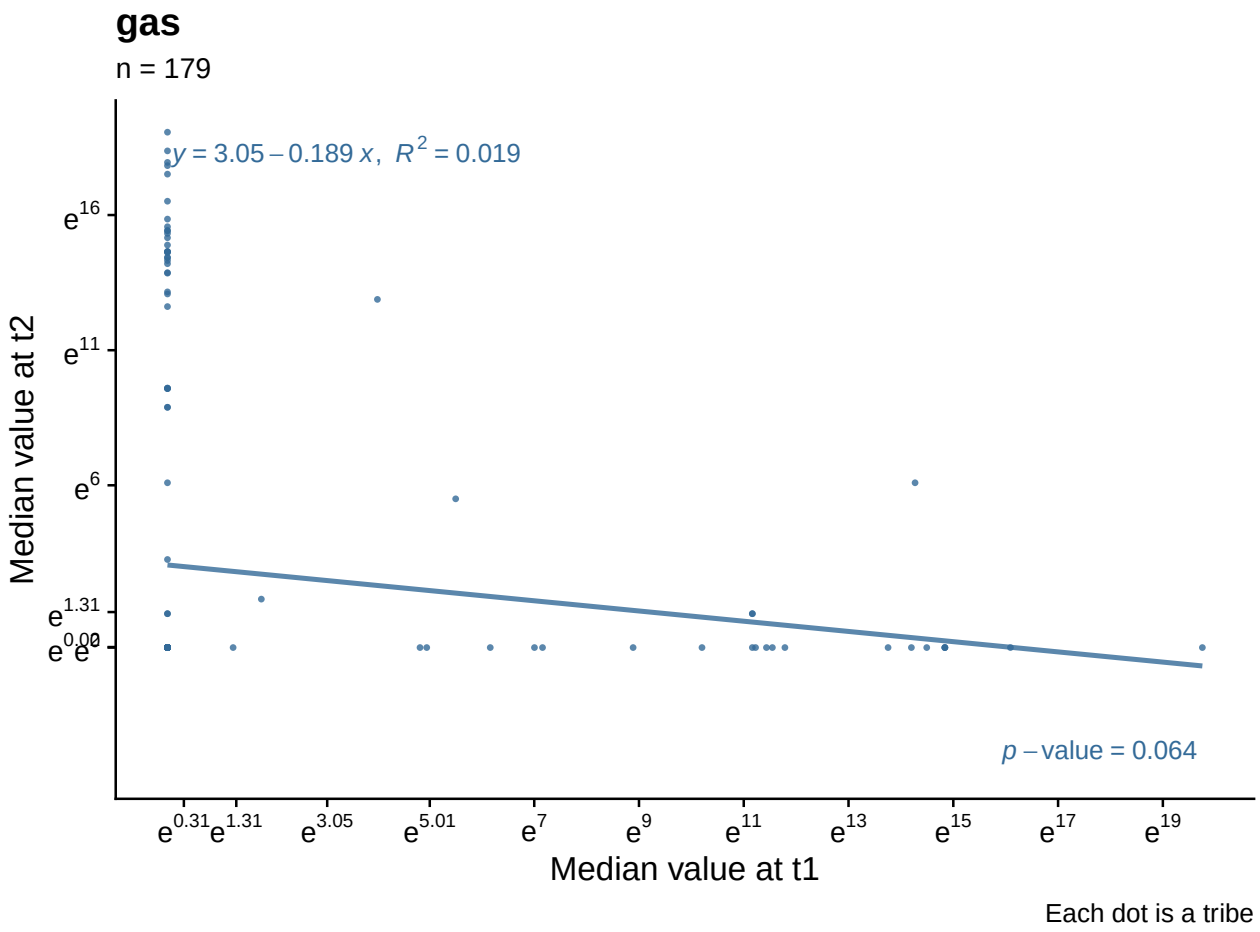
Q2. How are t1 and t2 related?

We first calculated the mean and median gas values of each tribe at t1 and t2 respectively. The reason why we also looked at the median is because the median is extreme-value resistent while the mean is heavily influence by outliers. We then made a scatterplot of the (log1p transformed) t2 means vs. t1 means, and another scatterplot of the (log1p transformed) t2 medians vs. t1 medians. These scatterplots showed weak positive linear relationships:

- **weak**: the more scattered the dots, the weaker the correlation.
- **positive**: upward slanted trend from bottom left corner to upper right corner. Or y tends to increase as x increases.

Finally, we ran linear regressions to quantify these relationships. For the log1p transformed mean values, we obtained a r-squared value of 0.043, which translates to a correlation of 0.21 (the squared root of 0.043), i.e., the correlation between the log1p transformed mean values at t1 and t2 is 0.21. The slope of the line is 0.203, meaning that for every 1000-unit (or 6.909-unit in log1p scale) increase in gas production at t1, we can expect a 3.07-unit (or 1.402-unit in log1p scale) increase at t2. This is statistically signicant as the p-value is $0.006 < 0.05$. A similar interpretation can be done for the median values.


gas
n = 179
y = 3.04 + 0.203 x, R2 = 0.043
p – value = 0.006
Mean value at t2
Mean value at t1
e19
e17
e15
e13
e11
e9
e7
e5.01
e3.05
e1.31
e0.31
Each dot is a tribe


gas
n = 179
y = 3.05 – 0.189 x, R2 = 0.019
p – value = 0.064
Median value at t2
Median value at t1
e16
e11
e6
e1.31
e0.00
e0.31
e1.31
e3.05
e5.01
e7
e9
e11
e13
e15
e17
e19
Each dot is a tribe