Modeling of Columbia, MO's Viral Load Sequenced in Wastewater Samples

The dataset referred to in this document is a result of the sampling and testing of wastewater collected from the Columbia, MO wastewater treatment plant. Samples were processed by a Mizzou lab led by Marc Johnson, with data processing conducted in collaboration with David O'Connor at University of Wisconsin. Additional clarification provided by David O'Connor includes:

- Each sample is deep sequenced, generating ~1 billion individual sequencing reads that we analyze. 'Total reads' is the number of sequencing reads in the sample.
- There is a fully documented and open workflow for processing the ~1 billion reads to determine what viruses are represented. The exact steps are documented here: https://github.com/dhoconno/nvd
- The final step in this workflow is determining how many of the total reads support a particular pathogen's presence - that number of reads is the 'mapped reads'

In the spirit of producing an analysis self-contained to this lab produced dataset – it's worth analyzing how viral concentration relates primarily to three variables: total reads, virus type, and sample week. Virus type categorizes genetic sequences into Respiratory, Veterinary, or Diarrheal viruses. Sample week represents when in the year these sample were taken, one through fifty-two. Total reads represent sequencing effort.

Total reads are specifically included as it is directly within the lab's control. Analyzing total reads provides insight into how sequencing depth affects measured virus concentrations, which in turn helps standardize and optimize laboratory processes. Including total reads as a predictor quantifies how sequencing effort impacts measurable detection of viral sequences. This brings us to an operational analysis of the way the overall dataset is collected, rather than a biological one. The question to answer is:

> "*Given a virus type, and time of year, how does adjusting sequencing efforts affect reported concentration?*"

Throughout the analysis, the terms Rate Per Billion & Parts Per Billion are used interchangeably.

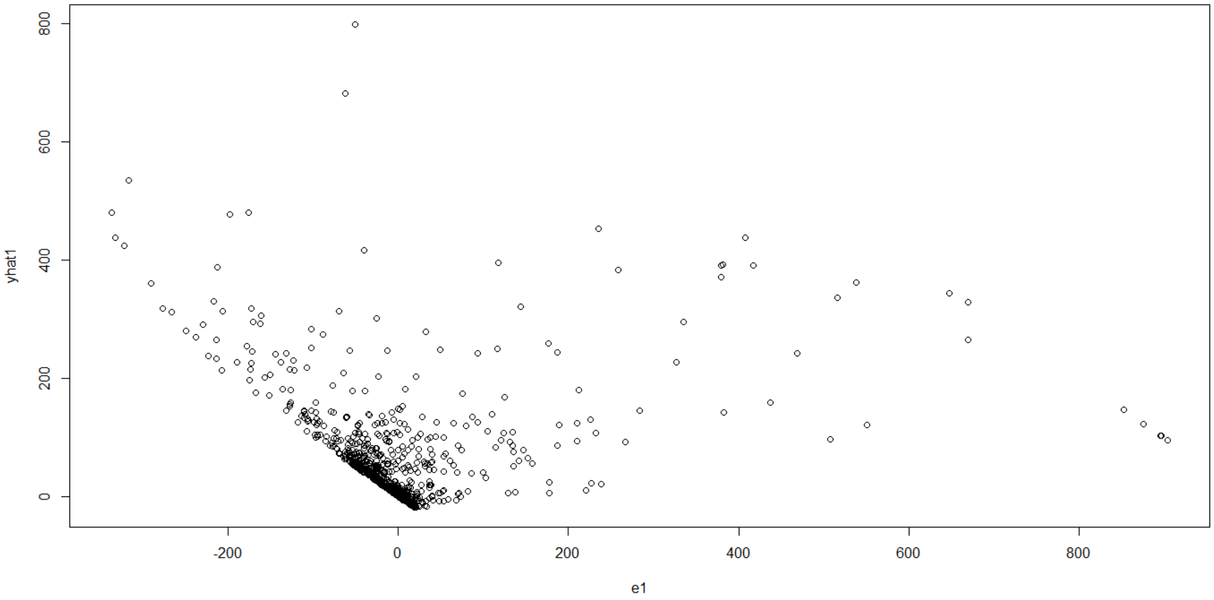
Fit Model 1: Full Model

```
lm(formula = partsperbillion ~ sampleweek + totalreads + virustype,
   data = virusdata)

Residuals:
   Min     1Q  Median     3Q    Max
-336.59  -28.83   -5.76    8.47  904.49

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.292e+01  8.681e+00   6.097 1.50e-09 ***
sampleweek         -5.402e-01  2.180e-01  -2.478   0.0134 *
totalreads          4.327e-10  2.217e-11  19.519  < 2e-16 ***
virustypeRespiratory -4.314e+01  7.746e+00  -5.570 3.22e-08 ***
virustypeVeterinary  -4.533e+01  4.512e+01  -1.005   0.3153
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 109.1 on 1081 degrees of freedom
Multiple R-squared:  0.3854,            Adjusted R-squared:  0.3832
F-statistic: 169.5 on 4 and 1081 DF,  p-value: < 2.2e-16
```



The plot above appears to have some concentration, with a noticeable linear pattern, and a wide dispersion for all points which do not follow either trend. There is a heavy cluster near the origin, this indicates that most predictions and residuals are close to zero. The overall model is statistically significant, the p-value is extremely small with a high F-statistic. All predictors prove to be useful except the Veterinary virus type. It is logically expected for totalreads to play a large role in predicting RPB, but the value of Respiratory types specifically indicates there is a biological reason for the low p-value. Variables with a low p-value provide clear insight as predictors. However, the fan shape of the residual plot presents concerns regarding reliability of C.I.s and p-values.

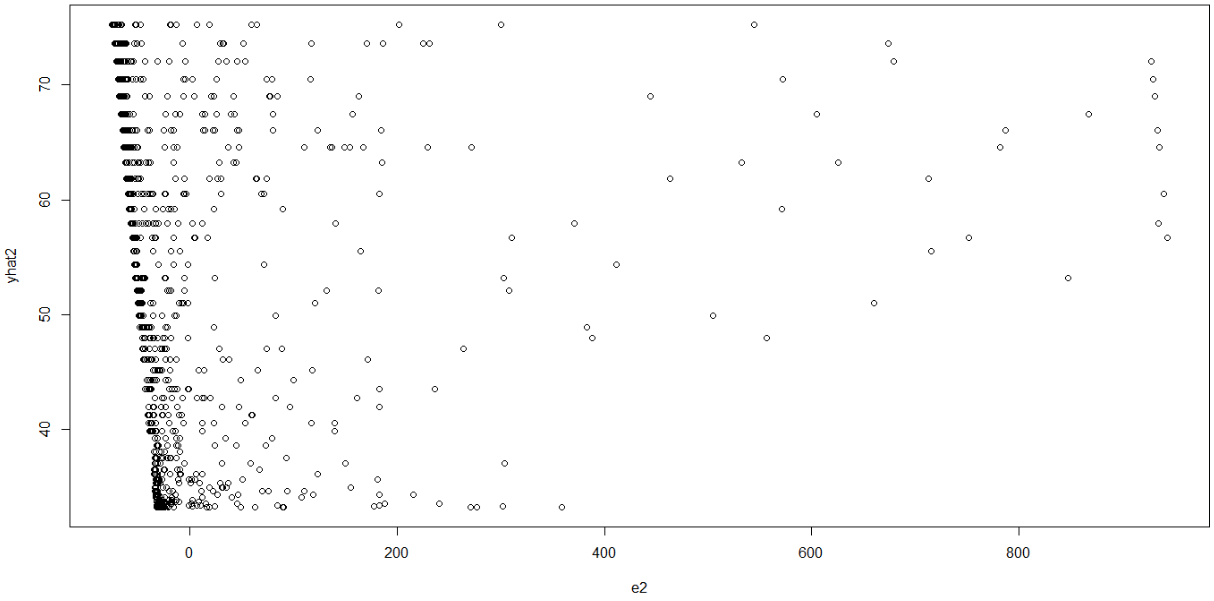
Fit Model 2: Quadratic

```
lm(formula = virusdata$partsperbillion ~ virusdata$sampleweek + 
    x1sq)

Residuals:
   Min     1Q Median     3Q    Max 
-74.89 -57.22 -34.80 -12.94 943.30 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           76.92641   10.33112   7.446 1.96e-13 ***
virusdata$sampleweek -1.66706    1.08461  -1.537    0.125    
x1sq                  0.01590    0.02097   0.758    0.449    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 138.4 on 1083 degrees of freedom
Multiple R-squared:  0.009968,	Adjusted R-squared:  0.00814 
F-statistic: 5.452 on 2 and 1083 DF,  p-value: 0.004405
```



The above model is just barely statistically significant. Neither variables appear to be significant in the model either, bringing us to the concerns around the residuals plot. The plot fans out quickly as values increase, though there is a general trend. R^2 is wildly low as well. With an R^2 value of 0.009968, less than 1% of variability in the response variable is explained by the model. It’s findings when digging into the details are negligible, and it’s hardly significant. Despite passing the F-test, I do not find it useful.

Fit Model 3: Choice

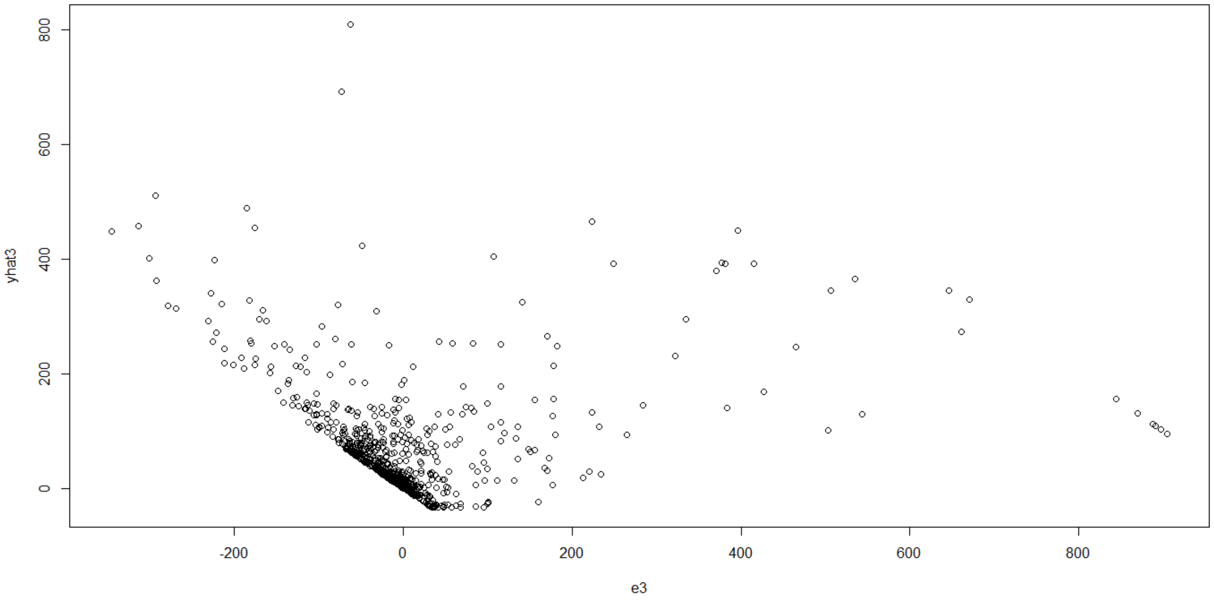
Results:

```
lm(formula = partsperbillion ~ totalreads + virustype + season,
   data = virusdata)

Residuals:
   Min      1Q  Median     3Q     Max
-344.68  -28.69  -10.08    8.31  904.44

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.667e+01  6.636e+00   5.525 4.13e-08 ***
totalreads           4.362e-10  2.192e-11  19.898  < 2e-16 ***
virustypeRespiratory -4.478e+01  7.686e+00  -5.826 7.47e-09 ***
virustypeVeterinary  -4.345e+01  4.484e+01  -0.969 0.332732
season.L            -2.393e+01  7.008e+00  -3.415 0.000662 ***
season.Q             1.625e+01  7.122e+00   2.282 0.022693 *
season.C             1.656e+01  7.299e+00   2.269 0.023462 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 108.5 on 1079 degrees of freedom
Multiple R-squared:  0.3943,	Adjusted R-squared:  0.3909
F-statistic: 117.1 on 6 and 1079 DF,  p-value: < 2.2e-16
```



A high F-statistic with 1079 degrees of freedom, and an extremely tiny p-value for the overall model makes model #3 statistically significant. Similar to the first, explained by the t-test all variables ***except*** the veterinary virus type are significant individually. Some advantages are the overall model significance (p-value: < 2.2e-16) & decent explanatory power, with roughly 39% of variability in RPB being explainable by the model. There is nuance captured in the impact of seasons. Again, the main disadvantage is visualized in the fan shape of the residual plot. This indicates the presence of heteroskedasticity, which just means the variance of residuals is not constant across the range of fitted values. The violation in assumptions is major and must be noted if this model is ever applied.

In summary, my preferred model is #3. It has a higher R^2 value to help in explaining variability in predictions and takes into account seasonality far better than model #1. The plot is ever so slightly tighter for #3. My highest concerns are the heteroskedasticity seen in all. However, these seem to be great outcomes for a preliminary analysis in model exploration.