STA 206 Final Project : Novozymes Enzyme Stability Prediction

Greg DePaul

Abstract

I explore regression, both multiple and nonlinear, on variable length finite state sequences. I explore these models with the goal of understanding how certain protein sequences imply their thermostability. This work also suggests the need for short term memory within any model architecture should anyone wish to accurately predict the thermostability of a sequence.

Introduction

The company Novozymes has released a Kaggle competition asking competitors to be able to predict enzyme thermostability. This work is valuable because the thermal stability of protein determines how much they can perform under harsh application conditions and/or their efficiency in serving as catalysts. This work is so valuable in fact that the top 3 scores win a cash prize, first place consisting of $12,000! The dataset itself only consists of the following varaibles:

- **Protein Sequence:** A variable length sequence of 20 possible states. An example sequence:

  `VPVNPEPDATSVENVALKTGSGDSQSDPIKADLEVKGQSALPFDVDCWAILCKGAPNVLQRVNEKTKNSNRDRSGANKGPFKDPQKWGIKALPPKNPSWSAQDFKS`

  Sequences can range from 200 to 30,000 states, which is why traditional models fail for this dataset. We need to be able to intake an adjustable amount of input.

- **pH:** The acidity level at which these sequences thermostability were measured. **Note:** There are some values in the dataset that exceed 14. These could be errors on the part of the recorder.

- **tm:** The measured thermostability. Response Variable.

The number of given points in this dataset is 313901 There is also an competition set with 2413 sequences.

So for this project, I will be trying to do the following:

1. Develop a model that will map a variable length sequence to a regressed value.
2. Approach whether or not a language model (with the concept of memory) is necessary.
3. Indicate connections that suggest stable or unstable protein sequences.

Methods and Results

Initial Data Exploration

For each of the given variables, we plot the histograms of frequency in the appendix. Specifically, the response variable we are trying to predict is highly clumped and centered around 48 degrees. It's likeliy this suggests a nonlinearity within our dataset. Therefore, we should avoid just jumping into a multiple linear regression. Instead, we turn towards to different models to try and capture this nonlinearity.

First Model

Given an enzyme sequence $(\{s_i\}_{i=1}^n, pH)$ pair, we consider important multivariable functions, similar to those we study in this class:

1. The **Predictive Value**, which should represent how much we trust an individual character sequence:

$$\text{predictive value}(s_i, pH) \sim \alpha_0 + \frac{\alpha_1}{\text{prevalence}(s_i)} + \frac{\alpha_2}{\text{variance}(s_i)} + \frac{\alpha_3}{1 + |pH - \text{median}\{pH(s_i)\}|}$$

2. The **Expected Value**, which represents what thermostability that sequence will likely take on based off that subset alone:

$$\text{expected value}(s_i) \sim \beta_0 + \beta_1 \text{mean}(s_i) + \beta_2 \max(s_i) + \beta_3 \min(s_i) + \beta_4 \text{median}(s_i) + \beta_5 \text{lower quartile}(s_i) + \beta_6 \text{upper quartile}(s_i)$$

We then construct a predictor model to be:

$$Y_{pred}(\{s_i\}_{i=1}^n, pH) = \frac{\sum_{i=1}^n \text{predictive value}(s_i, pH) \cdot \text{expected value}(s_i)}{\sum_{i=1}^n \text{predictive value}(s_i, pH)}$$

Our predictor function is nonlinear and therefore we turn to optimization to find our coefficient estimators. We define the loss function be

$$\mathcal{L}(Y_{pred}, Y_{true}) = \frac{1}{N} \sum_{k=1}^{N} (Y_{pred}^{(k)} - Y_{true}^{(k)})^2$$

where $N < 31390$. We can then optimize by minimizing this loss function over the few linear coefficients chosen above in order to minimize loss.

First Model Performance

Upon minimizing the mean square error over all of our labelled data, we obtained a MSE of 168 on train and 211 on our validation set. Submitting the competition dataset gives a Spearman correlation of 0.13. The current leading model has a score of 0.8, placing me at a rank of 1111 out of 1845 submitted models. From the figure labelled nonlinear model diagnostics, we see that our model lacks expressibility, specifically because of the large concentration of values that surround the mean of thermostability. This indicates that we are inferring on too few features. So instead we need to turn towards a more verbose model in order to solve this problem.

The difficulty with nonlinear models is interpreting the coefficients. So instead we shall turn to a more linear approach in order to gain a better understanding of the role these n-grams play.

Second Multilinear Model

We enumerate every 2-gram composed by any two states within the 20 possible achievable states for our sequences. We make the assumption that this model is NOT language in nature and therefore do not have a need for long term memory. We construct indicator functions of the following form:

$$\chi_B(\{s_i\}_{i=1}^n) = \begin{cases} 1 & \text{if } B \in \{s_i\}_{i=1}^n \\ 0 & \text{otherwise} \end{cases}$$

We then transform the dataset of tuples

$$(\{s_i\}_{i=1}^n, pH, tm) \to (pH, \underbrace{\chi_{jk}(\{s_i\}_{i=1}^n)}_{\text{400 features}}, tm)$$

Therefore, we have a model with 401 linear feature variables. This makes it very difficult to capture interactions because that would lead to $401^2 = 160801$ feature variables to consider. The only feature we can reasonably interact with is the ph level. Therefore, we are left considering the model:

$$f(tm) = \beta_0 + \beta_1 pH + \sum_{j=1}^{20}\sum_{k=1}^{20} \beta_{jk}\chi_{jk}(\{s_i\}_{i=1}^n) + pH\sum_{j=1}^{20}\sum_{k=1}^{20} \beta'_{jk}\chi_{jk}(\{s_i\}_{i=1}^n)$$

we allow f to be the function derived from the box-cox procedure. Specifically, we find a good value for lambda to be:

$$\lambda = 0.7474$$

which transforms our dataset and (hopefully) allows us to capture any nonlinearities. Therefore, we get for our model

$$\frac{1}{\lambda}\left((tm + shift)^\lambda - 1\right) = \beta_0 + \beta_1 pH + \sum_{j=1}^{20}\sum_{k=1}^{20} \beta_{jk}\chi_{jk}(\{s_i\}_{i=1}^n) + pH\sum_{j=1}^{20}\sum_{k=1}^{20} \beta'_{jk}\chi_{jk}(\{s_i\}_{i=1}^n)$$

Note: There is no possible way to explicitly write out this function. There are far too many feature variables. Please see the appendix for the model summary.

Linear Categorical Model Performance

We get an MSE of about 15.82 with an

$$R_a^2 \approx 0.41,$$

which is relatively good for a model of this size. Taking a look at the residuals versus fitted values, nothing suggests nonlinearity, as well as a pretty well behaved Q-Q plot, with the except of a couple of outliers.

We turn to the validation set to provide more contest. We see that the

$$MSPE_v = 16.99 < 13.82 \approx \frac{SSE}{N}$$

which suggests that there is no severe overfitting of our model and that it is potentially generalizable. We also see that the

$$R_a^2 \approx 0.50$$

, higher than that of the trained model! If we were to look at the estimated coefficients, we would see much of them are the same between the training model and the validation model, with the exceptions left for less significant estimators.

Lastly, we can look at the overlay of the predicted values on the validation set versus the true values. We see that compared to the previous nonlinear model, these distributions appear more similar in shape and value coverage. This suggests that the model does reasonably have the ability to achieve the expected values for thermostability.

Valuable Questions

- **How much of this data is explainable at this level?** Compared to other competitors using large scale language models like BERT, both models considered are incredibly tiny. But the current top place model is large BERT model with a Spearman coefficient score of 0.8, while this model yields a max score of 0.137. It's likely that this dataset does necessitate the concept of memory in order to predict values accurately.
- **Are some chain subsets more important than others?** Creating a model in this way, similar to those we have made all quarter long, allows us to analyze our coefficients and draw conclusions about the data our models describe. Observe, from the table listed under **Linear Categorical Model**, we can target very specific problematic sequence connections, such as 'NQ' and 'CR'. These coefficients also suggest potential replacements, such as 'YM' or 'MH'. We can tell from the summary table that these are very signifcant variables, with a p-value of being zero less than a significance level of 0.001. However, upon multiple choices of a random seed, these variables are subject to shift for different train / validation sets. Therefore, one should consider range of top coefficients as good and a range of bottom coefficients as bad instead of focusing specifically on the bottom or topmost coefficients.

  Also looking at these coefficients, we can see less useful interactions between pH and these connections make up the coefficients closest to zero. Checking their significance, they all fail to reject the hypothesis that their coefficients are anything but zero. HOWEVER, modelling without these interactions results in higher MSE. So for the moment, it is best to leave these variables within the model.

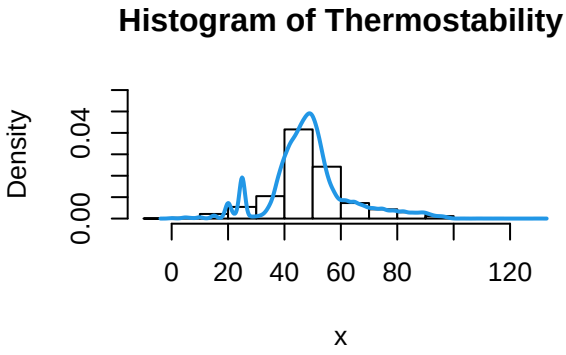
Conclusions and Limitations

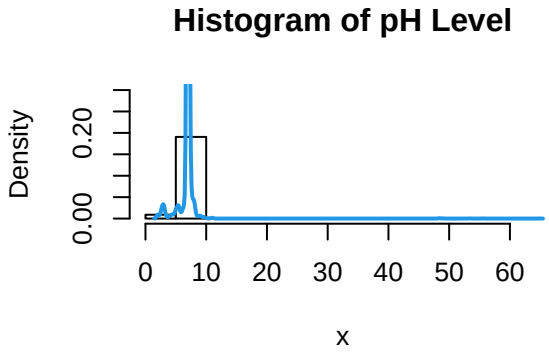
Multple regression models are able to provide a humanly interpretable way of understanding data, especially data as complicated as variable length sequence, such as DNA, Proteins, etc. Specifically, a categorical regression model is able to achieve a decent score in order to estimate this dataset. However, it's likely there are more nonlinearities at foot. OR, there are significant interactions that cannot be captured due to R's inability to handle large feature sets. These limitations mean only so many models can be approached.

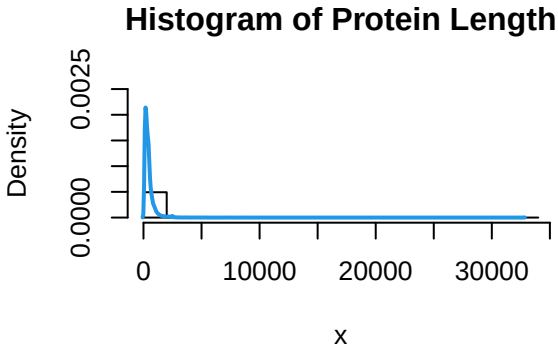
Appendix 1: Figures and Tables

Data Exploration

```
## [1] "tm quartiles:"

##    0%   25%   50%   75%  100%
##  -1.0  42.1  48.0  53.8 130.0

## [1] "pH quartiles:"

##    0%   25%   50%   75%  100%
##  1.99  7.00  7.00  7.00 64.90

## [1] "Protein Sequence Length quartiles:"

##      0%     25%     50%     75%    100%
##     5.0   197.0   335.5   523.0 32767.0
```

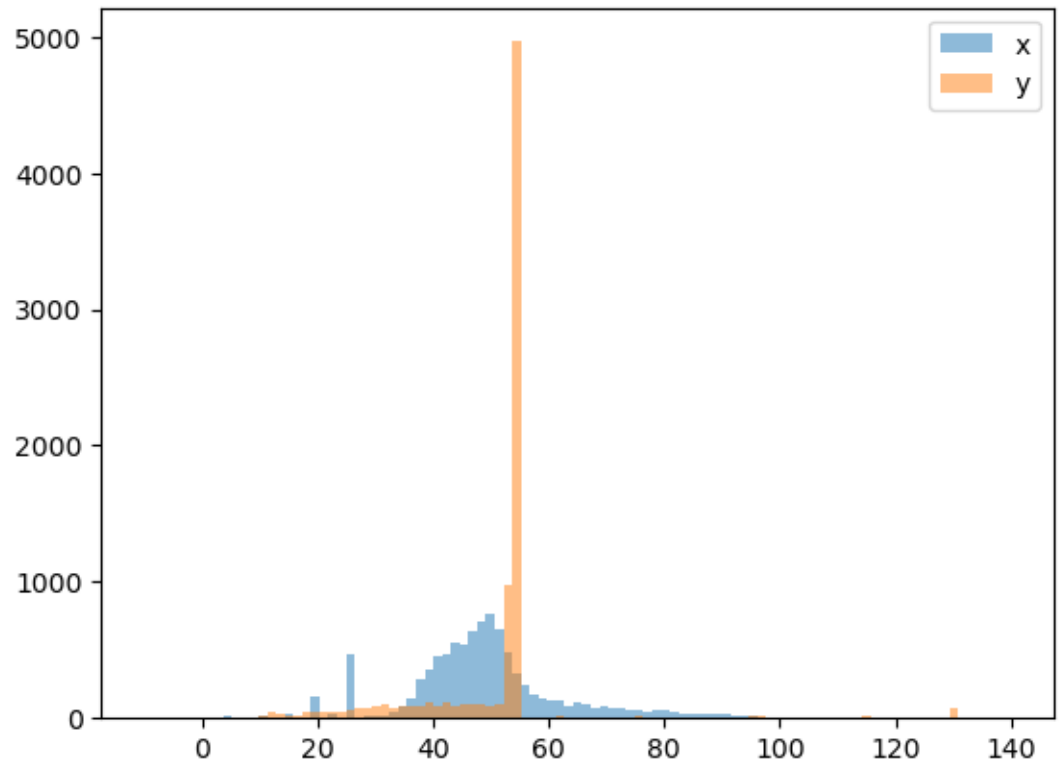




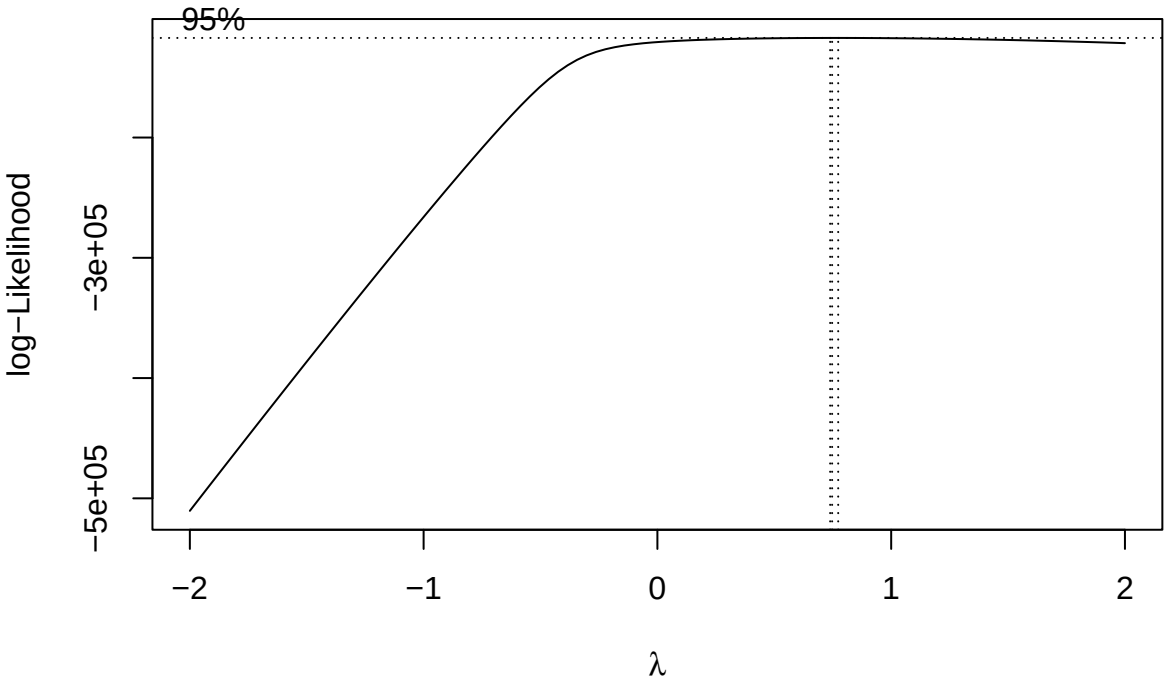


Nonlinear Model Diagnostics

```
knitr::include_graphics("nonlinear.png")
```



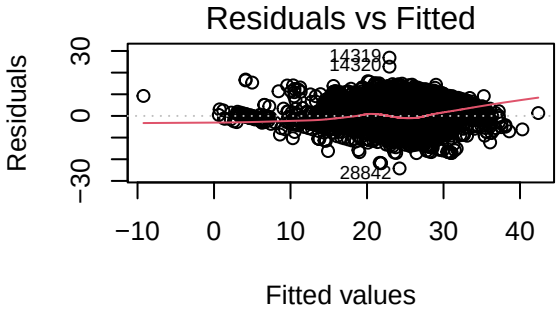
Linear Categorical Model

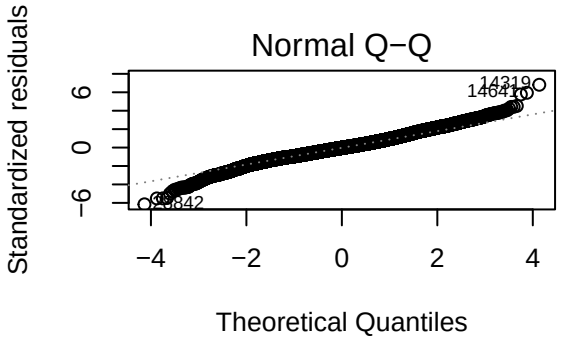


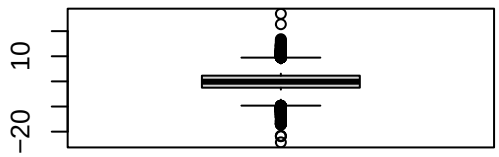
```
## [1] "lambda"

## [1] 0.7474747

## [1] "Maximum Coefficients"

##          FM (Intercept)          LR          MH          YM
##    32.02533    33.37129    34.94532    50.78894    54.25239

## [1] "Minimum Coefficients"

##        NQ        QY        GH        LC        RC
## -21.42729 -18.67073 -18.47352 -18.28280 -17.68644

## [1] "Least Impactful Coefficients"

##       NS:pH       TK:pH       DC:pH       FQ:pH       CT:pH
## 0.001537589 0.003389883 0.004490268 0.006307758 0.006605938
```

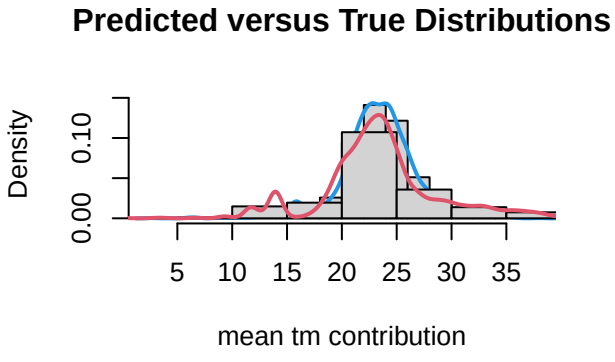
Linear Categorical Model Diagnostics

```
##                  SSE    R2_adj
## train_sum 433946.08 0.4178925
## valid_sum  33358.65 0.4875982

## [1] "MSPE"

## [1] 16.99149

## [1] "SSE / N"

## [1] 13.82434
```









Model Summary

```
##
## Call:
## lm(formula = Y_transform(tm) ~ . + pH:., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2471  -2.5021  -0.1275   2.2866  26.9008
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.371290   7.698657   4.335 1.46e-05 ***
## AA            0.251842   0.265967   0.947 0.343701
## AC            1.673137   2.941811   0.569 0.569535
## AD           -0.738618   2.281402  -0.324 0.746125
## AE           -0.207444   2.757807  -0.075 0.940040
## AF           -1.318022   1.709110  -0.771 0.440610
## AG            4.153387   1.905871   2.179 0.029321 *
## AH            0.601225   2.852147   0.211 0.833047
## AI            2.310886   1.811866   1.275 0.202172
## AK           -1.203998   2.954134  -0.408 0.683597
## AL            2.762523   2.247116   1.229 0.218946
## AM           -3.171273   2.226926  -1.424 0.154441
## AN           -4.055472   1.502670  -2.699 0.006962 **
## AP            4.063394   2.496424   1.628 0.103603
## AQ           -4.427142   2.111858  -2.096 0.036062 *
## AR           -5.536687   1.929585  -2.869 0.004116 **
## AS            3.166075   1.946524   1.627 0.103849
## AT           -0.572104   1.797310  -0.318 0.750251
## AV            3.356604   1.937153   1.733 0.083151 .
## AW           -3.819692   3.948444  -0.967 0.333357
## AY           -2.293320   2.047345  -1.120 0.262663
## CA            7.538234   2.612279   2.886 0.003909 **
## CC            4.503140   3.797965   1.186 0.235762
## CD           -0.832710   4.027399  -0.207 0.836198
## CE           -5.898752   3.165166  -1.864 0.062382 .
## CF           -1.550221   3.161679  -0.490 0.623914
## CG            4.875338   2.864198   1.702 0.088736 .
## CH            6.490394   3.955561   1.641 0.100845
## CI            6.739049   3.506434   1.922 0.054627 .
## CK            0.639356   5.243026   0.122 0.902944
## CL           -5.102740   3.965453  -1.287 0.198175
## CM           12.024472   3.777200   3.183 0.001457 **
## CN            1.627881   3.567742   0.456 0.648194
## CP           -6.798570   5.125504  -1.326 0.184712
## CQ            2.254214   4.801485   0.469 0.638728
## CR          -11.162951   5.387729  -2.072 0.038282 *
## CS            4.576979   2.793516   1.638 0.101344
## CT           -0.082852   4.538652  -0.018 0.985436
## CV            3.973194   3.129756   1.269 0.204277
## CW            3.560486   4.887264   0.729 0.466299
## CY           -4.242556   8.029430  -0.528 0.597243
## DA           -2.543688   2.108168  -1.207 0.227602
## DC           -0.203872   5.029419  -0.041 0.967666
## DD          -10.083442   5.737815  -1.757 0.078867 .
## DE            1.476504   3.474095   0.425 0.670837
## DF           -5.911865   2.838645  -2.083 0.037294 *
## DG           -2.682216   2.166620  -1.238 0.215737
## DH           -3.696027   4.194978  -0.881 0.378293
```

```
## DI          -10.783170   3.433517  -3.141 0.001688 **
## DK           -1.934470   2.382772  -0.812 0.416881
## DL           -0.470286   2.605627  -0.180 0.856770
## DM           -2.595117   4.941047  -0.525 0.599437
## DN            2.571835   3.360080   0.765 0.444035
## DP            0.022306   2.567672   0.009 0.993069
## DQ           -9.516072   3.364985  -2.828 0.004688 **
## DR            6.525448   2.546635   2.562 0.010401 *
## DS           -2.163460   2.490305  -0.869 0.384990
## DT          -10.481795   2.429281  -4.315 1.60e-05 ***
## DV            1.940118   1.899581   1.021 0.307102
## DW            6.202183   4.501223   1.378 0.168249
## DY           -2.833951   3.345447  -0.847 0.396943
## EA          -12.269121   3.182952  -3.855 0.000116 ***
## EC          -15.467792   6.931882  -2.231 0.025663 *
## ED           -3.505801   3.286948  -1.067 0.286170
## EE            1.210316   3.063376   0.395 0.692778
## EF            4.177517   3.152433   1.325 0.185125
## EG           -5.839717   3.855106  -1.515 0.129834
## EH            6.842141   5.845388   1.171 0.241802
## EI           -4.787500   4.028848  -1.188 0.234724
## EK          -12.668613   4.463442  -2.838 0.004539 **
## EL           -1.774635   2.160469  -0.821 0.411419
## EM           -4.396673   3.376994  -1.302 0.192945
## EN           -0.696048   2.602703  -0.267 0.789138
## EP           -3.808616   5.333780  -0.714 0.475199
## EQ           14.373682   6.953375   2.067 0.038729 *
## ER            0.274802   4.607194   0.060 0.952438
## ES           -0.490131   2.443083  -0.201 0.840997
## ET           -0.814397   2.431078  -0.335 0.737632
## EV            0.623793   0.804112   0.776 0.437901
## EW           -3.364192   3.350469  -1.004 0.315341
## EY           -3.395962   5.137262  -0.661 0.508589
## FA            0.965251   1.873012   0.515 0.606315
## FC           -6.009811   6.681011  -0.900 0.368375
## FD           10.961305   3.710765   2.954 0.003140 **
## FE            4.027422   3.558845   1.132 0.257785
## FF           -2.069867   2.703092  -0.766 0.443837
## FG            6.612574   2.963892   2.231 0.025686 *
## FH            7.023150   4.863982   1.444 0.148776
## FI           -0.505649   3.551861  -0.142 0.886795
## FK            1.037839   3.125542   0.332 0.739853
## FL            7.668867   3.238894   2.368 0.017904 *
## FM           32.025332   8.161312   3.924 8.73e-05 ***
## FN            1.975775   2.361407   0.837 0.402772
## FP           10.343792   4.078664   2.536 0.011216 *
## FQ            0.046685   2.276866   0.021 0.983641
## FR           -1.897173   2.206651  -0.860 0.389933
## FS           13.517476   5.140814   2.629 0.008557 **
## FT            0.068438   3.592201   0.019 0.984800
## FV           -7.308673   2.754351  -2.654 0.007971 **
## FW          -10.094144   6.511699  -1.550 0.121116
## FY           -2.877505   3.471590  -0.829 0.407184
## GA            0.613334   1.735526   0.353 0.723792
## GC           -2.784501   3.136714  -0.888 0.374703
## GD           -4.522103   2.661339  -1.699 0.089296 .
## GE           -5.883900   2.597168  -2.266 0.023489 *
## GF           -0.144900   2.155590  -0.067 0.946406
## GG            1.259330   2.405550   0.524 0.600624
## GH          -18.473520   4.581745  -4.032 5.55e-05 ***
```

```
## GI          -3.325580   3.851545  -0.863 0.387903
## GK          -2.635347   2.573856  -1.024 0.305896
## GL          -0.261826   3.402083  -0.077 0.938656
## GM          -8.187250   3.106866  -2.635 0.008413 **
## GN           6.146223   2.129677   2.886 0.003905 **
## GP          -4.298400   4.912579  -0.875 0.381593
## GQ           7.792210   3.994243   1.951 0.051084 .
## GR           8.482230   1.748310   4.852 1.23e-06 ***
## GS          -3.354082   3.708136  -0.905 0.365728
## GT           1.714112   2.377046   0.721 0.470848
## GV           0.157772   2.172421   0.073 0.942105
## GW           5.966027   4.327714   1.379 0.168041
## GY           2.235562   4.551495   0.491 0.623309
## HA          -4.246492   3.488494  -1.217 0.223506
## HC          -6.606594   5.566882  -1.187 0.235330
## HD           6.533986   5.133621   1.273 0.203106
## HE          -2.796770   4.001855  -0.699 0.484640
## HF          -1.318018   7.423564  -0.178 0.859081
## HG           8.869574   4.357470   2.035 0.041811 *
## HH          14.542776   8.033592   1.810 0.070269 .
## HI          -1.657348   3.535321  -0.469 0.639218
## HK           8.990528   5.598985   1.606 0.108342
## HL          -7.692089   3.749877  -2.051 0.040248 *
## HM          -3.665963   7.231586  -0.507 0.612203
## HN           2.774822   6.702666   0.414 0.678886
## HP           1.815154   4.997533   0.363 0.716451
## HQ          12.734438   5.407233   2.355 0.018526 *
## HR           7.920875   4.924984   1.608 0.107780
## HS          -6.090387   4.344354  -1.402 0.160954
## HT          -2.085799   4.432365  -0.471 0.637942
## HV           1.828256   4.421854   0.413 0.679273
## HW           0.225571   5.128185   0.044 0.964916
## HY         -15.388063   5.110471  -3.011 0.002606 **
## IA           4.935375   1.700269   2.903 0.003703 **
## IC           1.442463   4.182309   0.345 0.730175
## ID          -0.384623   4.087535  -0.094 0.925033
## IE          -5.911853   3.048896  -1.939 0.052510 .
## IF           2.077433   1.619438   1.283 0.199569
## IG           3.277363   3.893064   0.842 0.399881
## IH          -1.431814   6.566627  -0.218 0.827396
## II           1.322437   1.937921   0.682 0.494992
## IK          -4.295204   3.909945  -1.099 0.271981
## IL           0.668003   2.538089   0.263 0.792405
## IM          -4.313602   5.232327  -0.824 0.409712
## IN           0.653023   2.234243   0.292 0.770075
## IP         -11.792790   4.363677  -2.702 0.006886 **
## IQ          12.873446   5.452298   2.361 0.018228 *
## IR           0.985834   1.896304   0.520 0.603158
## IS           0.518756   2.382627   0.218 0.827645
## IT           5.119435   3.188085   1.606 0.108329
## IV          -0.181846   1.090160  -0.167 0.867524
## IW           2.176287   4.133378   0.527 0.598534
## IY           5.573189   3.343460   1.667 0.095547 .
## KA          -3.590331   2.093698  -1.715 0.086388 .
## KC          -1.469051   3.515290  -0.418 0.676021
## KD           3.986302   3.137221   1.271 0.203865
## KE           1.682458   3.330987   0.505 0.613498
## KF          -3.206174   2.883263  -1.112 0.266150
## KG           2.314413   1.983277   1.167 0.243235
## KH           2.422450   3.539358   0.684 0.493708
```

```
## KI          -6.806288   2.757890  -2.468 0.013596 *
## KK          -1.379541   2.699460  -0.511 0.609325
## KL           3.394528   2.818999   1.204 0.228538
## KM          -5.998454   3.299622  -1.818 0.069087 .
## KN          -1.689338   3.397549  -0.497 0.619036
## KP           4.329908   3.232536   1.339 0.180427
## KQ         -10.756612   4.057147  -2.651 0.008023 **
## KR          12.730959   4.273145   2.979 0.002892 **
## KS           8.593396   2.857462   3.007 0.002638 **
## KT          -4.265570   2.853055  -1.495 0.134903
## KV           0.986126   1.658336   0.595 0.552084
## KW          -7.007464   3.287937  -2.131 0.033076 *
## KY          -1.415132   2.521228  -0.561 0.574606
## LA           3.483964   2.214777   1.573 0.115718
## LC         -18.282795   4.593659  -3.980 6.91e-05 ***
## LD           6.193035   4.884249   1.268 0.204823
## LE           0.292900   2.647732   0.111 0.911916
## LF          -0.492163   2.086384  -0.236 0.813518
## LG          -0.466602   3.363467  -0.139 0.889667
## LH         -13.540319   6.030094  -2.245 0.024747 *
## LI          -0.746380   1.936926  -0.385 0.699987
## LK         -11.340922   4.035449  -2.810 0.004953 **
## LL          15.208879   3.419472   4.448 8.71e-06 ***
## LM           9.085976   3.675056   2.472 0.013429 *
## LN           0.672685   2.399227   0.280 0.779191
## LP           1.054806   5.242652   0.201 0.840546
## LQ          -5.175687   2.911871  -1.777 0.075506 .
## LR          34.945321   6.205415   5.631 1.80e-08 ***
## LS           2.303536   3.117607   0.739 0.459986
## LT          -1.273876   2.716417  -0.469 0.639106
## LV           9.919581   2.358822   4.205 2.62e-05 ***
## LW           3.030212   4.130722   0.734 0.463211
## LY          -9.886499   4.596167  -2.151 0.031482 *
## MA          -1.796350   2.291110  -0.784 0.433016
## MC         -15.703428   8.323845  -1.887 0.059230 .
## MD          -4.468059   3.596846  -1.242 0.214168
## ME          -3.770733   3.019042  -1.249 0.211682
## MF           2.255037   2.953960   0.763 0.445235
## MG          13.372589   3.917848   3.413 0.000643 ***
## MH          50.788937   9.209226   5.515 3.52e-08 ***
## MI           8.625043   2.981688   2.893 0.003823 **
## MK          -9.445996   3.439099  -2.747 0.006025 **
## ML           3.415158   5.163263   0.661 0.508339
## MM         -12.487315   4.406920  -2.834 0.004606 **
## MN           4.052843   2.905497   1.395 0.163061
## MP         -17.637056   7.649043  -2.306 0.021130 *
## MQ         -10.980301   4.734415  -2.319 0.020389 *
## MR           1.968732   4.883027   0.403 0.686820
## MS          26.039893   5.196883   5.011 5.46e-07 ***
## MT          -3.123872   3.934333  -0.794 0.427201
## MV           2.723155   3.035129   0.897 0.369614
## MW          -7.541345  11.486826  -0.657 0.511494
## MY           0.756325   5.956696   0.127 0.898965
## NA.         -5.341243   3.077504  -1.736 0.082650 .
## NC         -11.065060   6.445056  -1.717 0.086022 .
## ND           2.079693   2.717552   0.765 0.444110
## NE          -5.258260   2.674424  -1.966 0.049294 *
## NF           1.682282   2.329732   0.722 0.470244
## NG           1.528891   2.006635   0.762 0.446116
## NH           2.650914   5.931924   0.447 0.654958
```

```
## NI          -2.165802   1.582491  -1.369 0.171135
## NK           3.150250   2.784086   1.132 0.257846
## NL           4.975917   5.452337   0.913 0.361450
## NM           3.346354   3.591465   0.932 0.351473
## NN           6.733059   2.296945   2.931 0.003378 **
## NP           4.538017   2.076333   2.186 0.028854 *
## NQ         -21.427290   6.896823  -3.107 0.001893 **
## NR           8.151484   2.998571   2.718 0.006563 **
## NS          -0.033593   2.454878  -0.014 0.989082
## NT          -0.343965   2.334679  -0.147 0.882874
## NV           4.829978   1.705358   2.832 0.004626 **
## NW          -1.834333   3.767467  -0.487 0.626342
## NY          -6.849610   3.569304  -1.919 0.054991 .
## PA          -1.572497   2.115598  -0.743 0.457314
## PC           7.473498   6.318073   1.183 0.236868
## PD           0.382926   0.802235   0.477 0.633135
## PE           0.709066   3.679471   0.193 0.847189
## PF          -5.489079   3.659301  -1.500 0.133617
## PG           5.369650   2.727709   1.969 0.049014 *
## PH           0.097046   4.770922   0.020 0.983771
## PI          -4.854808   4.259532  -1.140 0.254400
## PK          -0.873602   1.759848  -0.496 0.619611
## PL           0.771579   5.160585   0.150 0.881149
## PM         -11.592805   3.446371  -3.364 0.000770 ***
## PN           0.204529   3.081811   0.066 0.947087
## PP          -3.945860   6.702856  -0.589 0.556079
## PQ          11.916629   4.879733   2.442 0.014610 *
## PR          -0.748728   6.912737  -0.108 0.913749
## PS          -4.077863   2.753495  -1.481 0.138624
## PT          -0.283772   5.161758  -0.055 0.956158
## PV          -1.021135   3.103374  -0.329 0.742128
## PW           9.124124   6.362682   1.434 0.151582
## PY          -5.034802   4.508838  -1.117 0.264153
## QA           8.562414   1.684620   5.083 3.75e-07 ***
## QC         -10.231699  10.212686  -1.002 0.316419
## QD         -10.079863   4.407079  -2.287 0.022192 *
## QE           4.544974   3.487869   1.303 0.192558
## QF          -4.633240   3.693909  -1.254 0.209747
## QG           4.226769   3.159237   1.338 0.180937
## QH         -15.484363   4.898428  -3.161 0.001574 **
## QI           2.896962   2.347095   1.234 0.217111
## QK           8.203032   3.839791   2.136 0.032662 *
## QL           4.299169   2.799401   1.536 0.124612
## QM          -1.645787   3.176968  -0.518 0.604437
## QN          -3.443400   1.803252  -1.910 0.056202 .
## QP          -2.051151   7.552084  -0.272 0.785931
## QQ          -3.762588   4.200204  -0.896 0.370362
## QR           0.974036   5.215195   0.187 0.851843
## QS          16.785894   3.995638   4.201 2.67e-05 ***
## QT           4.345539   3.044577   1.427 0.153503
## QV           5.496255   2.689760   2.043 0.041022 *
## QW          -3.384071   6.243867  -0.542 0.587835
## QY         -18.670730   3.886708  -4.804 1.57e-06 ***
## RA           1.381431   2.730519   0.506 0.612915
## RC         -17.686442   6.032995  -2.932 0.003375 **
## RD         -12.777364   5.574426  -2.292 0.021905 *
## RE           4.862699   3.365764   1.445 0.148539
## RF           3.223194   4.837071   0.666 0.505192
## RG           2.976831   1.806435   1.648 0.099384 .
## RH          -4.134227   5.398665  -0.766 0.443810
```

```
## RI          -6.314538   3.481496  -1.814 0.069728 .
## RK          -5.150882   3.177433  -1.621 0.105011
## RL         -10.380359   4.578511  -2.267 0.023386 *
## RM          -2.473807   2.221112  -1.114 0.265388
## RN           3.208119   3.569368   0.899 0.368771
## RP         -15.535863   6.616430  -2.348 0.018878 *
## RQ           0.521068   4.841174   0.108 0.914288
## RR          12.735082   3.955082   3.220 0.001284 **
## RS          -6.084212   3.099226  -1.963 0.049640 *
## RT          -0.266722   2.799151  -0.095 0.924088
## RV           2.058309   1.989420   1.035 0.300852
## RW          -7.068427   6.466059  -1.093 0.274334
## RY           4.641550   4.198885   1.105 0.268985
## SA          -5.335305   1.981423  -2.693 0.007093 **
## SC          -5.326562   3.515420  -1.515 0.129733
## SD           9.543527   5.501838   1.735 0.082822 .
## SE          -5.978904   1.493533  -4.003 6.27e-05 ***
## SF          -6.725628   2.851705  -2.358 0.018358 *
## SG          -8.605978   2.368611  -3.633 0.000280 ***
## SH         -14.526483   5.387000  -2.697 0.007010 **
## SI         -13.512250   3.427903  -3.942 8.11e-05 ***
## SK           0.101750   2.179315   0.047 0.962762
## SL           5.437006   3.977243   1.367 0.171628
## SM         -14.683910   5.382698  -2.728 0.006376 **
## SN         -15.333402   3.676330  -4.171 3.04e-05 ***
## SP           8.972976   3.866379   2.321 0.020307 *
## SQ           0.787392   3.608804   0.218 0.827285
## SR         -13.898194   3.645123  -3.813 0.000138 ***
## SS          -4.978179   4.856711  -1.025 0.305367
## ST           5.521297   2.457996   2.246 0.024695 *
## SV          -5.389093   2.436311  -2.212 0.026976 *
## SW          12.756801   4.574453   2.789 0.005296 **
## SY           2.654565   5.646732   0.470 0.638283
## TA          -1.426651   1.695301  -0.842 0.400057
## TC           6.353092   4.734900   1.342 0.179685
## TD           4.195214   2.349590   1.786 0.074190 .
## TE           2.314553   2.934594   0.789 0.430287
## TF          -1.910577   2.356044  -0.811 0.417415
## TG          -0.155120   3.411702  -0.045 0.963735
## TH          -1.579220   4.192585  -0.377 0.706422
## TI           4.327198   2.234040   1.937 0.052763 .
## TK          -0.165188   4.420825  -0.037 0.970193
## TL           9.376627   2.428895   3.860 0.000113 ***
## TM          -1.870251   5.185982  -0.361 0.718375
## TN          -2.650650   2.718489  -0.975 0.329546
## TP           1.360357   2.716864   0.501 0.616580
## TQ          -5.250716   5.517902  -0.952 0.341319
## TR           5.769279   2.861292   2.016 0.043776 *
## TS          -1.804645   2.684303  -0.672 0.501401
## TT          -7.436540   2.447343  -3.039 0.002379 **
## TV           3.240583   2.932691   1.105 0.269175
## TW           1.819156   3.123976   0.582 0.560355
## TY          -1.816605   3.536849  -0.514 0.607520
## VA           2.550544   2.865749   0.890 0.373469
## VC          -6.793598   3.266343  -2.080 0.037546 *
## VD           2.154551   2.266742   0.951 0.341864
## VE           3.545135   3.742517   0.947 0.343515
## VF          -1.314442   1.946365  -0.675 0.499471
## VG           0.436642   0.845509   0.516 0.605562
## VH          12.800090   5.771908   2.218 0.026587 *
```

```
## VI          -3.037511   2.315357  -1.312 0.189566
## VK          -8.200892   2.865551  -2.862 0.004214 **
## VL          10.396695   2.646655   3.928 8.58e-05 ***
## VM           5.879685   4.398623   1.337 0.181328
## VN          -3.611116   1.229709  -2.937 0.003321 **
## VP           1.828779   2.710301   0.675 0.499839
## VQ          -4.023329   3.638260  -1.106 0.268806
## VR           0.426917   0.825040   0.517 0.604846
## VS          -4.115033   2.809755  -1.465 0.143055
## VT          -0.688639   2.110095  -0.326 0.744159
## VV           0.716567   1.583734   0.452 0.650945
## VW          -0.203989   3.365648  -0.061 0.951671
## VY           0.679345   2.424598   0.280 0.779335
## WA           0.299577   3.365974   0.089 0.929081
## WC          -0.380526   5.397500  -0.071 0.943796
## WD         -14.905722   4.557427  -3.271 0.001074 **
## WE           3.798354   3.319517   1.144 0.252530
## WF          -4.276130   3.984667  -1.073 0.283215
## WG           5.981641   4.012202   1.491 0.136009
## WH          10.256966   7.847674   1.307 0.191221
## WI           3.003511   4.639405   0.647 0.517384
## WK           1.653608   6.010317   0.275 0.783220
## WL           3.211263   3.677444   0.873 0.382544
## WM          18.268448  12.589851   1.451 0.146779
## WN          -2.138299   4.487873  -0.476 0.633749
## WP           5.721330   3.283428   1.742 0.081434 .
## WQ          12.661689   5.406591   2.342 0.019193 *
## WR          22.262525   4.298046   5.180 2.24e-07 ***
## WS         -14.978859   7.449847  -2.011 0.044375 *
## WT         -13.316504   8.624810  -1.544 0.122606
## WV          -4.470267   2.379443  -1.879 0.060296 .
## WW          12.278371   8.610810   1.426 0.153901
## WY           6.542474   3.704006   1.766 0.077353 .
## YA           2.287386   2.004132   1.141 0.253740
## YC          25.303853   8.435350   3.000 0.002705 **
## YD           2.230713   6.899160   0.323 0.746447
## YE          -4.339344   4.325901  -1.003 0.315818
## YF          10.892529   3.193684   3.411 0.000649 ***
## YG           1.912616   2.263815   0.845 0.398194
## YH         -11.896582   6.642983  -1.791 0.073328 .
## YI          -6.899739   3.930312  -1.756 0.079182 .
## YK           9.939440   3.709002   2.680 0.007371 **
## YL           2.133204   3.492441   0.611 0.541333
## YM          54.252395   7.952783   6.822 9.18e-12 ***
## YN          -0.301163   4.669738  -0.064 0.948579
## YP          -4.106753   4.582276  -0.896 0.370140
## YQ          -2.152116   4.955043  -0.434 0.664053
## YR           0.739771   2.743239   0.270 0.787416
## YS           4.774251   3.882941   1.230 0.218878
## YT         -12.209799   4.669154  -2.615 0.008928 **
## YV           0.052200   0.781697   0.067 0.946759
## YW           9.237480   8.718758   1.059 0.289384
## YY           0.946769   6.428518   0.147 0.882915
## pH          -1.381413   1.098686  -1.257 0.208644
## AA:pH       -0.033564   0.037717  -0.890 0.373533
## AC:pH       -0.254527   0.420275  -0.606 0.544772
## AD:pH        0.069691   0.326108   0.214 0.830778
## AE:pH        0.042558   0.394000   0.108 0.913984
## AF:pH        0.219998   0.244317   0.900 0.367884
## AG:pH       -0.619510   0.272282  -2.275 0.022898 *
```

```
## AH:pH        -0.060330   0.407396  -0.148 0.882275
## AI:pH        -0.311334   0.259298  -1.201 0.229887
## AK:pH         0.131434   0.422147   0.311 0.755540
## AL:pH        -0.370215   0.321399  -1.152 0.249378
## AM:pH         0.442471   0.318365   1.390 0.164594
## AN:pH         0.571243   0.214895   2.658 0.007860 **
## AP:pH        -0.580910   0.356872  -1.628 0.103582
## AQ:pH         0.584973   0.301957   1.937 0.052722 .
## AR:pH         0.852849   0.275822   3.092 0.001990 **
## AS:pH        -0.454867   0.278270  -1.635 0.102139
## AT:pH         0.092195   0.257181   0.358 0.719985
## AV:pH        -0.376006   0.276644  -1.359 0.174104
## AW:pH         0.588629   0.563683   1.044 0.296376
## AY:pH         0.364819   0.292481   1.247 0.212289
## CA:pH        -1.092063   0.373083  -2.927 0.003424 **
## CC:pH        -0.629590   0.542537  -1.160 0.245873
## CD:pH         0.108227   0.575343   0.188 0.850794
## CE:pH         0.835854   0.452040   1.849 0.064458 .
## CF:pH         0.200112   0.451362   0.443 0.657514
## CG:pH        -0.724633   0.409249  -1.771 0.076631 .
## CH:pH        -0.916080   0.565042  -1.621 0.104973
## CI:pH        -0.954362   0.501128  -1.904 0.056865 .
## CK:pH        -0.113774   0.748759  -0.152 0.879227
## CL:pH         0.662809   0.566398   1.170 0.241923
## CM:pH        -1.769944   0.539782  -3.279 0.001043 **
## CN:pH        -0.215083   0.509681  -0.422 0.673031
## CP:pH         0.997933   0.731926   1.363 0.172757
## CQ:pH        -0.308542   0.685790  -0.450 0.652781
## CR:pH         1.574736   0.769213   2.047 0.040648 *
## CS:pH        -0.645061   0.399112  -1.616 0.106054
## CT:pH         0.006606   0.648238   0.010 0.991869
## CV:pH        -0.585194   0.447045  -1.309 0.190536
## CW:pH        -0.515854   0.698540  -0.738 0.460233
## CY:pH         0.591461   1.146886   0.516 0.606061
## DA:pH         0.363676   0.301391   1.207 0.227575
## DC:pH         0.004490   0.718312   0.006 0.995012
## DD:pH         1.428227   0.819214   1.743 0.081273 .
## DE:pH        -0.209975   0.496272  -0.423 0.672222
## DF:pH         0.834066   0.405453   2.057 0.039684 *
## DG:pH         0.346980   0.309751   1.120 0.262642
## DH:pH         0.542389   0.599158   0.905 0.365340
## DI:pH         1.569219   0.490235   3.201 0.001371 **
## DK:pH         0.259745   0.340558   0.763 0.445646
## DL:pH         0.108215   0.372255   0.291 0.771282
## DM:pH         0.350795   0.705492   0.497 0.619028
## DN:pH        -0.405230   0.480054  -0.844 0.398602
## DP:pH         0.055815   0.366817   0.152 0.879062
## DQ:pH         1.326654   0.480613   2.760 0.005778 **
## DR:pH        -0.966770   0.363834  -2.657 0.007884 **
## DS:pH         0.272451   0.355800   0.766 0.443836
## DT:pH         1.457537   0.346980   4.201 2.67e-05 ***
## DV:pH        -0.293180   0.271558  -1.080 0.280321
## DW:pH        -0.903488   0.642756  -1.406 0.159840
## DY:pH         0.431253   0.477883   0.902 0.366840
## EA:pH         1.753506   0.454617   3.857 0.000115 ***
## EC:pH         2.205489   0.989824   2.228 0.025878 *
## ED:pH         0.474976   0.469324   1.012 0.311527
## EE:pH        -0.116752   0.437714  -0.267 0.789679
## EF:pH        -0.629739   0.450241  -1.399 0.161923
## EG:pH         0.831693   0.550318   1.511 0.130725
```

```
## EH:pH        -0.981092   0.834765  -1.175 0.239888
## EI:pH         0.704922   0.575317   1.225 0.220482
## EK:pH         1.808062   0.636656   2.840 0.004516 **
## EL:pH         0.309491   0.308619   1.003 0.315954
## EM:pH         0.622729   0.482331   1.291 0.196685
## EN:pH         0.096701   0.371858   0.260 0.794828
## EP:pH         0.565985   0.761651   0.743 0.457425
## EQ:pH        -2.168452   0.993104  -2.184 0.029007 *
## ER:pH         0.014302   0.658142   0.022 0.982663
## ES:pH         0.078993   0.349020   0.226 0.820947
## ET:pH         0.104663   0.347323   0.301 0.763157
## EV:pH        -0.026301   0.114287  -0.230 0.817993
## EW:pH         0.466869   0.478661   0.975 0.329389
## EY:pH         0.440966   0.733572   0.601 0.547764
## FA:pH        -0.122508   0.267659  -0.458 0.647171
## FC:pH         0.837696   0.954145   0.878 0.379976
## FD:pH        -1.595780   0.530175  -3.010 0.002616 **
## FE:pH        -0.597989   0.508349  -1.176 0.239471
## FF:pH         0.304094   0.386257   0.787 0.431122
## FG:pH        -0.939570   0.423447  -2.219 0.026504 *
## FH:pH        -0.994116   0.694762  -1.431 0.152479
## FI:pH         0.101567   0.507476   0.200 0.841372
## FK:pH        -0.198040   0.446384  -0.444 0.657296
## FL:pH        -1.065962   0.462523  -2.305 0.021193 *
## FM:pH        -4.577641   1.165667  -3.927 8.62e-05 ***
## FN:pH        -0.269245   0.337362  -0.798 0.424825
## FP:pH        -1.460100   0.582342  -2.507 0.012172 *
## FQ:pH         0.006308   0.325156   0.019 0.984523
## FR:pH         0.282520   0.315391   0.896 0.370380
## FS:pH        -1.957224   0.734006  -2.666 0.007669 **
## FT:pH        -0.050084   0.513116  -0.098 0.922244
## FV:pH         1.041643   0.393448   2.647 0.008114 **
## FW:pH         1.453443   0.930019   1.563 0.118109
## FY:pH         0.436580   0.495931   0.880 0.378692
## GA:pH        -0.097470   0.248483  -0.392 0.694869
## GC:pH         0.370079   0.448194   0.826 0.408975
## GD:pH         0.643227   0.380250   1.692 0.090736 .
## GE:pH         0.935368   0.370801   2.523 0.011656 *
## GF:pH         0.047140   0.308010   0.153 0.878362
## GG:pH        -0.185941   0.343770  -0.541 0.588589
## GH:pH         2.678635   0.654366   4.093 4.26e-05 ***
## GI:pH         0.498191   0.550070   0.906 0.365109
## GK:pH         0.407238   0.367717   1.107 0.268097
## GL:pH         0.062799   0.485873   0.129 0.897161
## GM:pH         1.174199   0.443847   2.646 0.008162 **
## GN:pH        -0.875836   0.304319  -2.878 0.004005 **
## GP:pH         0.658084   0.701566   0.938 0.348242
## GQ:pH        -1.124033   0.570553  -1.970 0.048840 *
## GR:pH        -1.207341   0.249676  -4.836 1.33e-06 ***
## GS:pH         0.500648   0.529577   0.945 0.344477
## GT:pH        -0.228684   0.339793  -0.673 0.500946
## GV:pH        -0.014859   0.310498  -0.048 0.961831
## GW:pH        -0.874998   0.618123  -1.416 0.156912
## GY:pH        -0.284890   0.650107  -0.438 0.661230
## HA:pH         0.608600   0.498298   1.221 0.221961
## HC:pH         0.952326   0.795250   1.198 0.231115
## HD:pH        -0.934708   0.733210  -1.275 0.202385
## HE:pH         0.401529   0.571557   0.703 0.482362
## HF:pH         0.176328   1.060131   0.166 0.867901
## HG:pH        -1.252554   0.622337  -2.013 0.044160 *
```

```
## HH:pH        -2.102734   1.147378  -1.833 0.066867 .
## HI:pH         0.241339   0.504943   0.478 0.632688
## HK:pH        -1.317081   0.799282  -1.648 0.099399 .
## HL:pH         1.077151   0.535598   2.011 0.044323 *
## HM:pH         0.521547   1.032970   0.505 0.613633
## HN:pH        -0.415342   0.957347  -0.434 0.664403
## HP:pH        -0.259043   0.713819  -0.363 0.716684
## HQ:pH        -1.819874   0.772129  -2.357 0.018432 *
## HR:pH        -1.128500   0.703391  -1.604 0.108644
## HS:pH         0.846293   0.620262   1.364 0.172449
## HT:pH         0.289304   0.633126   0.457 0.647714
## HV:pH        -0.252798   0.631564  -0.400 0.688959
## HW:pH        -0.051962   0.732416  -0.071 0.943442
## HY:pH         2.178090   0.729847   2.984 0.002845 **
## IA:pH        -0.707573   0.243100  -2.911 0.003610 **
## IC:pH        -0.229210   0.597093  -0.384 0.701073
## ID:pH        -0.007409   0.583640  -0.013 0.989872
## IE:pH         0.840913   0.435444   1.931 0.053474 .
## IF:pH        -0.339211   0.231641  -1.464 0.143101
## IG:pH        -0.462034   0.556158  -0.831 0.406116
## IH:pH         0.184030   0.937807   0.196 0.844428
## II:pH        -0.205168   0.277017  -0.741 0.458922
## IK:pH         0.607017   0.558476   1.087 0.277084
## IL:pH        -0.098772   0.362884  -0.272 0.785482
## IM:pH         0.622590   0.747167   0.833 0.404701
## IN:pH        -0.086789   0.319267  -0.272 0.785748
## IP:pH         1.719324   0.623164   2.759 0.005801 **
## IQ:pH        -1.849199   0.778524  -2.375 0.017543 *
## IR:pH        -0.150694   0.271154  -0.556 0.578385
## IS:pH        -0.129343   0.340303  -0.380 0.703888
## IT:pH        -0.748765   0.455362  -1.644 0.100120
## IV:pH         0.027920   0.155453   0.180 0.857466
## IW:pH        -0.328849   0.590435  -0.557 0.577559
## IY:pH        -0.787240   0.477694  -1.648 0.099364 .
## KA:pH         0.523935   0.299190   1.751 0.079926 .
## KC:pH         0.196011   0.502252   0.390 0.696344
## KD:pH        -0.528041   0.448131  -1.178 0.238680
## KE:pH        -0.220825   0.475677  -0.464 0.642484
## KF:pH         0.425081   0.411973   1.032 0.302166
## KG:pH        -0.305747   0.283446  -1.079 0.280741
## KH:pH        -0.364852   0.505421  -0.722 0.470376
## KI:pH         0.961566   0.393987   2.441 0.014669 *
## KK:pH         0.130999   0.385689   0.340 0.734123
## KL:pH        -0.557818   0.402886  -1.385 0.166200
## KM:pH         0.830645   0.471292   1.762 0.077998 .
## KN:pH         0.244392   0.485153   0.504 0.614447
## KP:pH        -0.663735   0.461728  -1.438 0.150586
## KQ:pH         1.491927   0.579489   2.575 0.010042 *
## KR:pH        -1.799641   0.610171  -2.949 0.003187 **
## KS:pH        -1.302045   0.408083  -3.191 0.001421 **
## KT:pH         0.618505   0.407500   1.518 0.129075
## KV:pH        -0.131239   0.236962  -0.554 0.579695
## KW:pH         0.993947   0.469815   2.116 0.034387 *
## KY:pH         0.207830   0.360283   0.577 0.564044
## LA:pH        -0.501090   0.316633  -1.583 0.113533
## LC:pH         2.585105   0.656032   3.941 8.15e-05 ***
## LD:pH        -0.846238   0.696848  -1.214 0.224613
## LE:pH         0.062139   0.378123   0.164 0.869469
## LF:pH         0.072677   0.298171   0.244 0.807431
## LG:pH         0.106644   0.480460   0.222 0.824345
```

```
## LH:pH         1.928874   0.861085   2.240 0.025096 *
## LI:pH         0.095005   0.276787   0.343 0.731419
## LK:pH         1.661111   0.576295   2.882 0.003950 **
## LL:pH        -2.126539   0.488381  -4.354 1.34e-05 ***
## LM:pH        -1.317464   0.524890  -2.510 0.012080 *
## LN:pH        -0.095620   0.342980  -0.279 0.780408
## LP:pH        -0.175014   0.748827  -0.234 0.815206
## LQ:pH         0.709391   0.416194   1.704 0.088304 .
## LR:pH        -4.934925   0.885877  -5.571 2.56e-08 ***
## LS:pH        -0.319439   0.445370  -0.717 0.473229
## LT:pH         0.200455   0.387637   0.517 0.605077
## LV:pH        -1.437472   0.336823  -4.268 1.98e-05 ***
## LW:pH        -0.407368   0.589968  -0.690 0.489890
## LY:pH         1.460973   0.656395   2.226 0.026039 *
## MA:pH         0.249130   0.327549   0.761 0.446908
## MC:pH         2.240857   1.188301   1.886 0.059337 .
## MD:pH         0.697882   0.513841   1.358 0.174422
## ME:pH         0.553465   0.431239   1.283 0.199352
## MF:pH        -0.352794   0.422130  -0.836 0.403304
## MG:pH        -1.900937   0.559657  -3.397 0.000683 ***
## MH:pH        -7.242959   1.315083  -5.508 3.67e-08 ***
## MI:pH        -1.244361   0.425963  -2.921 0.003489 **
## MK:pH         1.345591   0.491240   2.739 0.006163 **
## ML:pH        -0.518990   0.737198  -0.704 0.481436
## MM:pH         1.762286   0.629533   2.799 0.005124 **
## MN:pH        -0.546384   0.415098  -1.316 0.188092
## MP:pH         2.526357   1.092336   2.313 0.020741 *
## MQ:pH         1.580890   0.676209   2.338 0.019401 *
## MR:pH        -0.257296   0.697436  -0.369 0.712192
## MS:pH        -3.754228   0.742073  -5.059 4.24e-07 ***
## MT:pH         0.438837   0.561873   0.781 0.434794
## MV:pH        -0.371516   0.433491  -0.857 0.391434
## MW:pH         1.076403   1.640808   0.656 0.511817
## MY:pH        -0.114180   0.850922  -0.134 0.893258
## NA.:pH        0.774572   0.439621   1.762 0.078096 .
## NC:pH         1.599297   0.920572   1.737 0.082348 .
## ND:pH        -0.271598   0.388087  -0.700 0.484034
## NE:pH         0.752548   0.382044   1.970 0.048872 *
## NF:pH        -0.241110   0.332871  -0.724 0.468866
## NG:pH        -0.242840   0.286682  -0.847 0.396963
## NH:pH        -0.384957   0.847201  -0.454 0.649554
## NI:pH         0.294084   0.226431   1.299 0.194031
## NK:pH        -0.482651   0.397617  -1.214 0.224812
## NL:pH        -0.757379   0.778586  -0.973 0.330680
## NM:pH        -0.438685   0.513068  -0.855 0.392545
## NN:pH        -0.966555   0.328097  -2.946 0.003222 **
## NP:pH        -0.628295   0.296797  -2.117 0.034276 *
## NQ:pH         3.046056   0.984951   3.093 0.001986 **
## NR:pH        -1.185716   0.428320  -2.768 0.005639 **
## NS:pH         0.001538   0.350679   0.004 0.996502
## NT:pH         0.034999   0.333693   0.105 0.916470
## NV:pH        -0.685478   0.243948  -2.810 0.004959 **
## NW:pH         0.231540   0.538271   0.430 0.667086
## NY:pH         0.994937   0.509873   1.951 0.051026 .
## PA:pH         0.224895   0.302272   0.744 0.456874
## PC:pH        -1.046859   0.902418  -1.160 0.246035
## PD:pH        -0.050994   0.114203  -0.447 0.655227
## PE:pH        -0.084580   0.525493  -0.161 0.872131
## PF:pH         0.793114   0.522665   1.517 0.129167
## PG:pH        -0.762202   0.389774  -1.955 0.050534 .
```

```
## PH:pH         0.011611   0.681309   0.017 0.986403
## PI:pH         0.659387   0.608439   1.084 0.278491
## PK:pH         0.097813   0.251148   0.389 0.696935
## PL:pH        -0.064748   0.736735  -0.088 0.929969
## PM:pH         1.648834   0.492107   3.351 0.000808 ***
## PN:pH        -0.064864   0.440242  -0.147 0.882867
## PP:pH         0.602482   0.957332   0.629 0.529136
## PQ:pH        -1.663369   0.696854  -2.387 0.016995 *
## PR:pH         0.127026   0.987001   0.129 0.897597
## PS:pH         0.572237   0.393474   1.454 0.145869
## PT:pH         0.022420   0.737183   0.030 0.975738
## PV:pH         0.172601   0.443206   0.389 0.696955
## PW:pH        -1.282653   0.908619  -1.412 0.158064
## PY:pH         0.717248   0.644141   1.113 0.265506
## QA:pH        -1.225826   0.240855  -5.089 3.61e-07 ***
## QC:pH         1.436018   1.458582   0.985 0.324863
## QD:pH         1.450636   0.629331   2.305 0.021172 *
## QE:pH        -0.609340   0.498266  -1.223 0.221370
## QF:pH         0.606733   0.527449   1.150 0.250024
## QG:pH        -0.601364   0.451165  -1.333 0.182571
## QH:pH         2.173773   0.699732   3.107 0.001895 **
## QI:pH        -0.424447   0.335579  -1.265 0.205947
## QK:pH        -1.215767   0.548371  -2.217 0.026628 *
## QL:pH        -0.697059   0.399947  -1.743 0.081366 .
## QM:pH         0.207839   0.453881   0.458 0.647018
## QN:pH         0.474113   0.257601   1.840 0.065706 .
## QP:pH         0.241948   1.078433   0.224 0.822485
## QQ:pH         0.488639   0.599948   0.814 0.415383
## QR:pH        -0.154858   0.744832  -0.208 0.835301
## QS:pH        -2.444719   0.570632  -4.284 1.84e-05 ***
## QT:pH        -0.660857   0.434871  -1.520 0.128608
## QV:pH        -0.816151   0.384265  -2.124 0.033685 *
## QW:pH         0.455540   0.891548   0.511 0.609387
## QY:pH         2.659396   0.555011   4.792 1.66e-06 ***
## RA:pH        -0.125636   0.390220  -0.322 0.747483
## RC:pH         2.495509   0.861664   2.896 0.003781 **
## RD:pH         1.849903   0.795871   2.324 0.020113 *
## RE:pH        -0.638203   0.480449  -1.328 0.184075
## RF:pH        -0.446401   0.690752  -0.646 0.518120
## RG:pH        -0.396086   0.258113  -1.535 0.124906
## RH:pH         0.574748   0.771049   0.745 0.456030
## RI:pH         0.932802   0.497195   1.876 0.060648 .
## RK:pH         0.750432   0.453680   1.654 0.098119 .
## RL:pH         1.524450   0.653673   2.332 0.019701 *
## RM:pH         0.341049   0.317512   1.074 0.282775
## RN:pH        -0.474341   0.509724  -0.931 0.352077
## RP:pH         2.274497   0.944866   2.407 0.016081 *
## RQ:pH        -0.098711   0.691452  -0.143 0.886482
## RR:pH        -1.803155   0.564803  -3.193 0.001412 **
## RS:pH         0.863767   0.442700   1.951 0.051051 .
## RT:pH         0.037835   0.399913   0.095 0.924627
## RV:pH        -0.249133   0.284595  -0.875 0.381366
## RW:pH         1.023497   0.923400   1.108 0.267699
## RY:pH        -0.636730   0.599776  -1.062 0.288421
## SA:pH         0.733900   0.283414   2.590 0.009617 **
## SC:pH         0.761754   0.502080   1.517 0.129229
## SD:pH        -1.371127   0.785460  -1.746 0.080886 .
## SE:pH         0.840946   0.213750   3.934 8.37e-05 ***
## SF:pH         0.948647   0.407317   2.329 0.019866 *
## SG:pH         1.240784   0.338472   3.666 0.000247 ***
```

```
## SH:pH        2.042939   0.769171   2.656 0.007911 **
## SI:pH        1.942745   0.489699   3.967 7.29e-05 ***
## SK:pH       -0.061017   0.311333  -0.196 0.844622
## SL:pH       -0.752370   0.567935  -1.325 0.185266
## SM:pH        2.128466   0.768767   2.769 0.005632 **
## SN:pH        2.169941   0.525126   4.132 3.60e-05 ***
## SP:pH       -1.246972   0.552066  -2.259 0.023908 *
## SQ:pH       -0.139029   0.515376  -0.270 0.787345
## SR:pH        2.000057   0.520472   3.843 0.000122 ***
## SS:pH        0.580122   0.693255   0.837 0.402708
## ST:pH       -0.832890   0.350954  -2.373 0.017641 *
## SV:pH        0.736862   0.348151   2.117 0.034311 *
## SW:pH       -1.826021   0.653352  -2.795 0.005196 **
## SY:pH       -0.379104   0.806346  -0.470 0.638251
## TA:pH        0.205467   0.242298   0.848 0.396449
## TC:pH       -0.919937   0.676070  -1.361 0.173616
## TD:pH       -0.606978   0.335572  -1.809 0.070495 .
## TE:pH       -0.338498   0.419121  -0.808 0.419307
## TF:pH        0.258181   0.336705   0.767 0.443214
## TG:pH        0.031919   0.487319   0.065 0.947778
## TH:pH        0.232999   0.598834   0.389 0.697214
## TI:pH       -0.636569   0.319257  -1.994 0.046172 *
## TK:pH        0.003390   0.631340   0.005 0.995716
## TL:pH       -1.373048   0.346958  -3.957 7.60e-05 ***
## TM:pH        0.266988   0.740697   0.360 0.718509
## TN:pH        0.388812   0.388431   1.001 0.316845
## TP:pH       -0.183239   0.387936  -0.472 0.636686
## TQ:pH        0.737187   0.787850   0.936 0.349439
## TR:pH       -0.834131   0.408759  -2.041 0.041296 *
## TS:pH        0.251901   0.383374   0.657 0.511146
## TT:pH        1.059350   0.349473   3.031 0.002437 **
## TV:pH       -0.470229   0.419066  -1.122 0.261835
## TW:pH       -0.267112   0.446375  -0.598 0.549577
## TY:pH        0.246884   0.505287   0.489 0.625128
## VA:pH       -0.369616   0.409590  -0.902 0.366849
## VC:pH        0.954878   0.466654   2.046 0.040744 *
## VD:pH       -0.342046   0.323691  -1.057 0.290655
## VE:pH       -0.528724   0.534524  -0.989 0.322599
## VF:pH        0.198636   0.278109   0.714 0.475086
## VG:pH       -0.101249   0.120323  -0.841 0.400089
## VH:pH       -1.801312   0.824034  -2.186 0.028826 *
## VI:pH        0.433045   0.330650   1.310 0.190316
## VK:pH        1.170568   0.409319   2.860 0.004242 **
## VL:pH       -1.511800   0.377658  -4.003 6.27e-05 ***
## VM:pH       -0.827772   0.627966  -1.318 0.187454
## VN:pH        0.535172   0.176045   3.040 0.002368 **
## VP:pH       -0.266580   0.387145  -0.689 0.491094
## VQ:pH        0.589607   0.519574   1.135 0.256473
## VR:pH       -0.046937   0.117455  -0.400 0.689443
## VS:pH        0.597110   0.401134   1.489 0.136616
## VT:pH        0.095591   0.301316   0.317 0.751059
## VV:pH       -0.072630   0.226357  -0.321 0.748314
## VW:pH        0.040310   0.480878   0.084 0.933196
## VY:pH       -0.054189   0.346389  -0.156 0.875688
## WA:pH       -0.024908   0.480764  -0.052 0.958682
## WC:pH        0.085904   0.771040   0.111 0.911290
## WD:pH        2.152965   0.650983   3.307 0.000943 ***
## WE:pH       -0.518677   0.474127  -1.094 0.273981
## WF:pH        0.572990   0.569046   1.007 0.313976
## WG:pH       -0.830751   0.573040  -1.450 0.147146
```

```
## WH:pH        -1.434557   1.120985  -1.280 0.200651
## WI:pH        -0.433453   0.662553  -0.654 0.512978
## WK:pH        -0.252079   0.858446  -0.294 0.769031
## WL:pH        -0.447061   0.525308  -0.851 0.394751
## WM:pH        -2.629265   1.798286  -1.462 0.143727
## WN:pH         0.279918   0.641081   0.437 0.662380
## WP:pH        -0.842341   0.468677  -1.797 0.072303 .
## WQ:pH        -1.817005   0.772044  -2.353 0.018605 *
## WR:pH        -3.179187   0.613954  -5.178 2.26e-07 ***
## WS:pH         2.129949   1.063778   2.002 0.045268 *
## WT:pH         1.916067   1.231640   1.556 0.119790
## WV:pH         0.638590   0.340000   1.878 0.060364 .
## WW:pH        -1.741671   1.228923  -1.417 0.156426
## WY:pH        -0.940965   0.529193  -1.778 0.075396 .
## YA:pH        -0.329149   0.286507  -1.149 0.250634
## YC:pH        -3.629337   1.204347  -3.014 0.002585 **
## YD:pH        -0.288809   0.985176  -0.293 0.769406
## YE:pH         0.645483   0.617925   1.045 0.296219
## YF:pH        -1.563135   0.456213  -3.426 0.000613 ***
## YG:pH        -0.256631   0.323323  -0.794 0.427358
## YH:pH         1.694530   0.948877   1.786 0.074138 .
## YI:pH         1.030250   0.561453   1.835 0.066520 .
## YK:pH        -1.409984   0.529824  -2.661 0.007790 **
## YL:pH        -0.289859   0.498733  -0.581 0.561116
## YM:pH        -7.730628   1.135579  -6.808 1.01e-11 ***
## YN:pH         0.072722   0.666850   0.109 0.913161
## YP:pH         0.597393   0.654534   0.913 0.361408
## YQ:pH         0.296492   0.707635   0.419 0.675227
## YR:pH        -0.064110   0.391970  -0.164 0.870081
## YS:pH        -0.697754   0.554458  -1.258 0.208242
## YT:pH         1.732279   0.666713   2.598 0.009375 **
## YV:pH         0.015645   0.111291   0.141 0.888203
## YW:pH        -1.311643   1.244490  -1.054 0.291911
## YY:pH        -0.097231   0.917845  -0.106 0.915635
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.976 on 27449 degrees of freedom
## Multiple R-squared:  0.4344, Adjusted R-squared:  0.4179
## F-statistic: 26.32 on 801 and 27449 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
##
## Response: Y_transform(tm)
##               Df Sum Sq Mean Sq   F value    Pr(>F)
## AA             1     51    51.1    3.2293 0.0723415 .
## AC             1   1542  1542.0   97.5355 < 2.2e-16 ***
## AD             1   2630  2630.0  166.3604 < 2.2e-16 ***
## AE             1   4898  4898.0  309.8185 < 2.2e-16 ***
## AF             1    441   441.3   27.9173 1.276e-07 ***
## AG             1    190   190.2   12.0309 0.0005241 ***
## AH             1    936   936.0   59.2080 1.466e-14 ***
## AI             1   1065  1064.5   67.3365 2.388e-16 ***
## AK             1   4413  4413.0  279.1426 < 2.2e-16 ***
## AL             1    971   971.5   61.4490 4.708e-15 ***
## AM             1    242   241.7   15.2863 9.260e-05 ***
## AN             1   3056  3056.3  193.3259 < 2.2e-16 ***
## AP             1   3502  3501.9  221.5099 < 2.2e-16 ***
## AQ             1   6158  6158.0  389.5236 < 2.2e-16 ***
## AR             1  19741 19740.6 1248.6821 < 2.2e-16 ***
```

```
## AS            1    939   939.4   59.4185 1.318e-14 ***
## AT            1    300   300.1   18.9813 1.325e-05 ***
## AV            1  11547 11547.2  730.4085 < 2.2e-16 ***
## AW            1    289   289.5   18.3100 1.884e-05 ***
## AY            1     28    27.5    1.7405 0.1870823
## CA            1   1865  1865.2  117.9828 < 2.2e-16 ***
## CC            1    964   964.3   60.9941 5.928e-15 ***
## CD            1   1200  1199.9   75.8985 < 2.2e-16 ***
## CE            1   1511  1511.0   95.5791 < 2.2e-16 ***
## CF            1   4924  4923.9  311.4598 < 2.2e-16 ***
## CG            1    235   234.7   14.8486 0.0001168 ***
## CH            1    755   754.7   47.7390 4.975e-12 ***
## CI            1    464   463.7   29.3307 6.153e-08 ***
## CK            1    859   858.9   54.3285 1.744e-13 ***
## CL            1  10210 10210.3  645.8452 < 2.2e-16 ***
## CM            1   3388  3388.3  214.3251 < 2.2e-16 ***
## CN            1    327   327.2   20.6964 5.405e-06 ***
## CP            1    138   138.0    8.7277 0.0031367 **
## CQ            1    442   442.4   27.9852 1.232e-07 ***
## CR            1    207   207.1   13.1025 0.0002954 ***
## CS            1    199   198.6   12.5595 0.0003948 ***
## CT            1    295   295.5   18.6909 1.543e-05 ***
## CV            1    591   590.8   37.3719 9.893e-10 ***
## CW            1     15    14.7    0.9284 0.3352806
## CY            1     68    68.4    4.3264 0.0375348 *
## DA            1    262   261.8   16.5572 4.734e-05 ***
## DC            1      9     8.8    0.5561 0.4558343
## DD            1    508   507.7   32.1165 1.466e-08 ***
## DE            1     93    93.3    5.9003 0.0151450 *
## DF            1     18    17.7    1.1174 0.2904916
## DG            1   1833  1832.8  115.9346 < 2.2e-16 ***
## DH            1     27    26.9    1.7004 0.1922500
## DI            1    687   686.7   43.4375 4.456e-11 ***
## DK            1   3812  3811.9  241.1172 < 2.2e-16 ***
## DL            1   3074  3074.4  194.4687 < 2.2e-16 ***
## DM            1    211   211.3   13.3636 0.0002570 ***
## DN            1   4799  4798.7  303.5377 < 2.2e-16 ***
## DP            1   1793  1793.5  113.4459 < 2.2e-16 ***
## DQ            1    513   512.8   32.4397 1.242e-08 ***
## DR            1    364   364.4   23.0474 1.589e-06 ***
## DS            1   2775  2775.3  175.5487 < 2.2e-16 ***
## DT            1   2261  2261.1  143.0223 < 2.2e-16 ***
## DV            1    486   485.9   30.7374 2.981e-08 ***
## DW            1    192   191.6   12.1225 0.0004990 ***
## DY            1    603   603.3   38.1600 6.609e-10 ***
## EA            1     37    36.5    2.3097 0.1285818
## EC            1      2     1.6    0.1023 0.7491126
## ED            1     18    17.8    1.1290 0.2880032
## EE            1   3517  3517.1  222.4700 < 2.2e-16 ***
## EF            1   1258  1258.1   79.5791 < 2.2e-16 ***
## EG            1    194   193.9   12.2660 0.0004620 ***
## EH            1      0     0.1    0.0059 0.9387400
## EI            1     72    72.1    4.5604 0.0327290 *
## EK            1     38    37.7    2.3822 0.1227382
## EL            1   4567  4567.3  288.9031 < 2.2e-16 ***
## EM            1      2     2.0    0.1270 0.7215470
## EN            1    887   887.4   56.1338 6.974e-14 ***
## EP            1    136   136.3    8.6226 0.0033230 **
## EQ            1  12717 12717.2  804.4173 < 2.2e-16 ***
## ER            1   2589  2588.6  163.7393 < 2.2e-16 ***
```

```
## ES            1     43    43.1    2.7264 0.0987114 .
## ET            1     13    12.6    0.7996 0.3712122
## EV            1   2018  2017.6  127.6194 < 2.2e-16 ***
## EW            1     64    63.6    4.0204 0.0449641 *
## EY            1    665   665.4   42.0892 8.869e-11 ***
## FA            1    392   392.0   24.7933 6.420e-07 ***
## FC            1     47    46.6    2.9496 0.0859090 .
## FD            1    254   253.8   16.0548 6.170e-05 ***
## FE            1    561   561.3   35.5028 2.578e-09 ***
## FF            1    182   182.3   11.5323 0.0006849 ***
## FG            1    749   749.2   47.3872 5.952e-12 ***
## FH            1      0     0.4    0.0280 0.8670600
## FI            1     59    58.9    3.7285 0.0535032 .
## FK            1   2468  2468.1  156.1163 < 2.2e-16 ***
## FL            1    923   923.4   58.4088 2.199e-14 ***
## FM            1    329   329.3   20.8306 5.039e-06 ***
## FN            1     39    39.2    2.4786 0.1154188
## FP            1    434   433.5   27.4230 1.647e-07 ***
## FQ            1     23    23.3    1.4733 0.2248414
## FR            1    167   167.0   10.5647 0.0011541 **
## FS            1   1465  1465.2   92.6825 < 2.2e-16 ***
## FT            1   1552  1551.5   98.1403 < 2.2e-16 ***
## FV            1     11    11.4    0.7223 0.3954001
## FW            1    208   208.4   13.1821 0.0002831 ***
## FY            1   1018  1017.8   64.3831 1.065e-15 ***
## GA            1    269   268.6   16.9914 3.766e-05 ***
## GC            1      6     6.3    0.3986 0.5278366
## GD            1    360   360.0   22.7685 1.837e-06 ***
## GE            1   6528  6527.6  412.9009 < 2.2e-16 ***
## GF            1    602   602.4   38.1017 6.809e-10 ***
## GG            1    536   535.9   33.8958 5.879e-09 ***
## GH            1   1180  1180.3   74.6584 < 2.2e-16 ***
## GI            1     11    11.0    0.6935 0.4049672
## GK            1    419   418.8   26.4886 2.669e-07 ***
## GL            1   1174  1174.1   74.2673 < 2.2e-16 ***
## GM            1    191   191.1   12.0871 0.0005085 ***
## GN            1    170   169.5   10.7245 0.0010586 **
## GP            1    686   685.5   43.3638 4.627e-11 ***
## GQ            1    242   241.5   15.2781 9.301e-05 ***
## GR            1    418   417.8   26.4282 2.754e-07 ***
## GS            1    196   195.8   12.3881 0.0004328 ***
## GT            1     69    69.3    4.3807 0.0363568 *
## GV            1     77    76.9    4.8630 0.0274461 *
## GW            1     96    95.5    6.0434 0.0139644 *
## GY            1   1284  1284.3   81.2405 < 2.2e-16 ***
## HA            1     52    52.3    3.3104 0.0688521 .
## HC            1     37    36.7    2.3237 0.1274270
## HD            1    473   472.6   29.8920 4.608e-08 ***
## HE            1     34    33.7    2.1309 0.1443643
## HF            1     20    19.8    1.2504 0.2634861
## HG            1    714   713.9   45.1584 1.853e-11 ***
## HH            1    516   516.3   32.6558 1.111e-08 ***
## HI            1    159   159.2   10.0675 0.0015108 **
## HK            1   1569  1569.4   99.2715 < 2.2e-16 ***
## HL            1     93    93.2    5.8947 0.0151927 *
## HM            1     59    58.8    3.7169 0.0538736 .
## HN            1    561   560.6   35.4580 2.638e-09 ***
## HP            1    143   143.2    9.0580 0.0026179 **
## HQ            1     14    14.0    0.8838 0.3471734
## HR            1    421   420.8   26.6193 2.495e-07 ***
```

```
## HS            1    269   269.2   17.0252 3.699e-05 ***
## HT            1    176   175.7   11.1115 0.0008591 ***
## HV            1     58    58.4    3.6920 0.0546852 .
## HW            1    169   169.0   10.6929 0.0010768 **
## HY            1    822   821.8   51.9847 5.738e-13 ***
## IA            1     20    19.6    1.2390 0.2656643
## IC            1    281   281.0   17.7761 2.493e-05 ***
## ID            1   2114  2113.8  133.7063 < 2.2e-16 ***
## IE            1      0     0.0    0.0024 0.9612203
## IF            1   1471  1471.0   93.0451 < 2.2e-16 ***
## IG            1     85    85.3    5.3986 0.0201599 *
## IH            1      0     0.2    0.0122 0.9120495
## II            1    262   262.5   16.6028 4.621e-05 ***
## IK            1     14    13.6    0.8601 0.3537282
## IL            1     12    12.3    0.7810 0.3768444
## IM            1    143   143.0    9.0466 0.0026343 **
## IN            1     94    93.7    5.9280 0.0149087 *
## IP            1    889   888.5   56.2042 6.730e-14 ***
## IQ            1    633   633.2   40.0522 2.511e-10 ***
## IR            1    182   182.0   11.5103 0.0006931 ***
## IS            1   3317  3317.2  209.8263 < 2.2e-16 ***
## IT            1    406   405.7   25.6627 4.092e-07 ***
## IV            1     51    51.0    3.2269 0.0724468 .
## IW            1    344   344.1   21.7680 3.091e-06 ***
## IY            1     17    16.6    1.0531 0.3048021
## KA            1     25    24.7    1.5595 0.2117518
## KC            1     46    46.2    2.9246 0.0872481 .
## KD            1    543   543.5   34.3777 4.591e-09 ***
## KE            1     36    36.2    2.2928 0.1299844
## KF            1    119   118.6    7.5029 0.0061640 **
## KG            1     96    96.3    6.0896 0.0136042 *
## KH            1    254   253.9   16.0633 6.142e-05 ***
## KI            1    215   215.0   13.5968 0.0002270 ***
## KK            1   1585  1584.9  100.2543 < 2.2e-16 ***
## KL            1   2047  2046.8  129.4713 < 2.2e-16 ***
## KM            1    131   130.7    8.2688 0.0040363 **
## KN            1    467   467.0   29.5384 5.528e-08 ***
## KP            1   1179  1179.0   74.5760 < 2.2e-16 ***
## KQ            1   1767  1767.3  111.7893 < 2.2e-16 ***
## KR            1     28    27.7    1.7539 0.1853959
## KS            1   2685  2685.1  169.8472 < 2.2e-16 ***
## KT            1      6     6.0    0.3812 0.5369695
## KV            1      0     0.1    0.0033 0.9544784
## KW            1     21    21.4    1.3564 0.2441667
## KY            1     41    41.0    2.5906 0.1075117
## LA            1      6     6.3    0.3972 0.5285287
## LC            1     34    33.7    2.1328 0.1441899
## LD            1    216   215.5   13.6336 0.0002226 ***
## LE            1   2808  2807.9  177.6127 < 2.2e-16 ***
## LF            1     52    52.1    3.2970 0.0694161 .
## LG            1    788   788.1   49.8514 1.698e-12 ***
## LH            1      7     7.3    0.4613 0.4970189
## LI            1    256   255.8   16.1826 5.767e-05 ***
## LK            1    260   260.0   16.4458 5.020e-05 ***
## LL            1    893   893.2   56.4963 5.802e-14 ***
## LM            1    138   138.5    8.7577 0.0030856 **
## LN            1    186   186.5   11.7959 0.0005945 ***
## LP            1    193   193.1   12.2176 0.0004742 ***
## LQ            1   1029  1029.3   65.1082 7.377e-16 ***
## LR            1   1593  1593.1  100.7698 < 2.2e-16 ***
```

```
## LS            1     74    73.5    4.6523 0.0310207 *
## LT            1      3     3.0    0.1913 0.6618209
## LV            1      3     2.8    0.1772 0.6737612
## LW            1    526   525.7   33.2526 8.180e-09 ***
## LY            1    581   581.4   36.7773 1.342e-09 ***
## MA            1      5     5.2    0.3269 0.5675242
## MC            1      0     0.0    0.0024 0.9612085
## MD            1   1258  1257.9   79.5666 < 2.2e-16 ***
## ME            1      5     5.5    0.3476 0.5554795
## MF            1     47    47.0    2.9709 0.0847868 .
## MG            1      1     0.7    0.0465 0.8292936
## MH            1     39    38.5    2.4363 0.1185678
## MI            1     30    30.5    1.9262 0.1651912
## MK            1    130   130.5    8.2546 0.0040681 **
## ML            1   1154  1153.6   72.9684 < 2.2e-16 ***
## MM            1    110   109.6    6.9329 0.0084672 **
## MN            1    845   844.9   53.4460 2.731e-13 ***
## MP            1     51    51.3    3.2433 0.0717256 .
## MQ            1      0     0.1    0.0067 0.9347578
## MR            1    328   328.4   20.7717 5.196e-06 ***
## MS            1   1298  1298.2   82.1181 < 2.2e-16 ***
## MT            1     75    75.3    4.7656 0.0290416 *
## MV            1    138   137.6    8.7021 0.0031812 **
## MW            1    148   148.0    9.3606 0.0022192 **
## MY            1     21    21.5    1.3570 0.2440764
## NA.           1     19    19.4    1.2285 0.2677091
## NC            1     26    26.2    1.6561 0.1981394
## ND            1    243   243.1   15.3794 8.815e-05 ***
## NE            1     59    59.4    3.7586 0.0525474 .
## NF            1     32    31.9    2.0160 0.1556598
## NG            1     99    98.5    6.2321 0.0125513 *
## NH            1     29    29.3    1.8502 0.1737711
## NI            1    153   153.1    9.6832 0.0018616 **
## NK            1     98    98.3    6.2192 0.0126430 *
## NL            1    983   983.2   62.1944 3.227e-15 ***
## NM            1     75    74.9    4.7382 0.0295092 *
## NN            1     12    12.2    0.7729 0.3793412
## NP            1    195   194.5   12.3052 0.0004524 ***
## NQ            1     31    30.6    1.9371 0.1639976
## NR            1    400   399.6   25.2741 5.005e-07 ***
## NS            1    324   324.3   20.5135 5.946e-06 ***
## NT            1    390   390.4   24.6968 6.750e-07 ***
## NV            1      4     3.9    0.2474 0.6188963
## NW            1    212   212.4   13.4321 0.0002478 ***
## NY            1     19    19.2    1.2160 0.2701550
## PA            1     49    49.5    3.1299 0.0768810 .
## PC            1     85    85.4    5.3993 0.0201526 *
## PD            1    144   144.0    9.1078 0.0025476 **
## PE            1    201   200.9   12.7078 0.0003647 ***
## PF            1    103   102.9    6.5107 0.0107279 *
## PG            1     25    25.1    1.5863 0.2078594
## PH            1    416   416.2   26.3280 2.900e-07 ***
## PI            1    198   197.6   12.4966 0.0004084 ***
## PK            1    540   540.3   34.1745 5.095e-09 ***
## PL            1    912   911.8   57.6744 3.191e-14 ***
## PM            1      0     0.3    0.0172 0.8957484
## PN            1    321   321.4   20.3279 6.551e-06 ***
## PP            1    485   485.1   30.6863 3.061e-08 ***
## PQ            1     37    36.7    2.3212 0.1276319
## PR            1    507   507.2   32.0832 1.492e-08 ***
```

```
## PS            1    134   134.0    8.4780 0.0035975 **
## PT            1    517   517.0   32.6997 1.087e-08 ***
## PV            1    339   338.7   21.4235 3.699e-06 ***
## PW            1     96    96.2    6.0838 0.0136490 *
## PY            1    209   208.7   13.2007 0.0002804 ***
## QA            1     12    11.9    0.7555 0.3847480
## QC            1     69    69.4    4.3898 0.0361635 *
## QD            1    217   217.4   13.7515 0.0002091 ***
## QE            1    237   237.2   15.0044 0.0001075 ***
## QF            1    603   603.1   38.1492 6.646e-10 ***
## QG            1      6     6.1    0.3862 0.5343233
## QH            1    397   396.5   25.0823 5.528e-07 ***
## QI            1      9     9.0    0.5706 0.4500321
## QK            1    405   404.9   25.6091 4.207e-07 ***
## QL            1   1276  1275.6   80.6847 < 2.2e-16 ***
## QM            1    291   290.8   18.3945 1.802e-05 ***
## QN            1    314   314.4   19.8888 8.241e-06 ***
## QP            1    947   946.7   59.8821 1.042e-14 ***
## QQ            1    512   512.0   32.3880 1.275e-08 ***
## QR            1     16    15.9    1.0035 0.3164642
## QS            1    855   854.7   54.0650 1.994e-13 ***
## QT            1   1070  1069.5   67.6508 < 2.2e-16 ***
## QV            1    525   525.0   33.2060 8.378e-09 ***
## QW            1      3     3.0    0.1875 0.6650228
## QY            1     12    11.5    0.7300 0.3929016
## RA            1    704   703.6   44.5083 2.581e-11 ***
## RC            1    188   187.6   11.8666 0.0005724 ***
## RD            1    221   221.1   13.9837 0.0001848 ***
## RE            1    710   709.7   44.8910 2.123e-11 ***
## RF            1    323   323.3   20.4508 6.144e-06 ***
## RG            1    322   321.6   20.3429 6.500e-06 ***
## RH            1    145   145.1    9.1766 0.0024536 **
## RI            1     26    26.0    1.6461 0.1994949
## RK            1    112   111.7    7.0630 0.0078739 **
## RL            1    866   865.7   54.7608 1.400e-13 ***
## RM            1      5     4.9    0.3105 0.5773767
## RN            1    237   237.2   15.0068 0.0001074 ***
## RP            1    477   476.7   30.1560 4.022e-08 ***
## RQ            1     88    87.7    5.5484 0.0185043 *
## RR            1     42    42.3    2.6762 0.1018699
## RS            1      5     5.2    0.3299 0.5657301
## RT            1      3     3.3    0.2107 0.6461852
## RV            1    194   193.9   12.2642 0.0004625 ***
## RW            1      1     0.6    0.0353 0.8509076
## RY            1    404   403.9   25.5467 4.346e-07 ***
## SA            1    118   118.0    7.4644 0.0062970 **
## SC            1     10    10.4    0.6555 0.4181715
## SD            1      5     4.9    0.3131 0.5757986
## SE            1      7     7.0    0.4412 0.5065388
## SF            1     48    47.6    3.0088 0.0828252 .
## SG            1     12    12.3    0.7782 0.3777138
## SH            1    323   322.7   20.4108 6.274e-06 ***
## SI            1    108   108.4    6.8572 0.0088333 **
## SK            1    650   649.6   41.0908 1.477e-10 ***
## SL            1      6     6.4    0.4080 0.5230052
## SM            1    180   180.4   11.4131 0.0007303 ***
## SN            1     45    45.4    2.8702 0.0902453 .
## SP            1     74    74.3    4.7006 0.0301604 *
## SQ            1    502   502.0   31.7547 1.766e-08 ***
## SR            1    204   204.3   12.9208 0.0003255 ***
```

```
## SS            1   2481  2481.3  156.9527 < 2.2e-16 ***
## ST            1    258   257.6   16.2921 5.444e-05 ***
## SV            1    465   464.9   29.4073 5.915e-08 ***
## SW            1     42    41.6    2.6331 0.1046705
## SY            1      4     3.9    0.2485 0.6181145
## TA            1    152   152.4    9.6408 0.0019050 **
## TC            1     51    51.4    3.2535 0.0712801 .
## TD            1      3     2.5    0.1612 0.6880356
## TE            1      1     1.5    0.0924 0.7611884
## TF            1     16    16.1    1.0208 0.3123307
## TG            1     65    64.7    4.0934 0.0430610 *
## TH            1      3     2.8    0.1787 0.6725294
## TI            1    263   262.7   16.6168 4.587e-05 ***
## TK            1    124   123.7    7.8275 0.0051494 **
## TL            1    213   212.9   13.4682 0.0002431 ***
## TM            1     27    27.5    1.7367 0.1875657
## TN            1     92    91.6    5.7948 0.0160798 *
## TP            1      0     0.1    0.0087 0.9257755
## TQ            1     16    16.0    1.0151 0.3136838
## TR            1     21    21.3    1.3500 0.2452875
## TS            1      0     0.0    0.0019 0.9648963
## TT            1     31    30.6    1.9361 0.1641087
## TV            1     42    42.2    2.6689 0.1023407
## TW            1    204   203.9   12.8966 0.0003298 ***
## TY            1      2     1.6    0.0989 0.7531825
## VA            1      0     0.1    0.0050 0.9438112
## VC            1      6     5.7    0.3594 0.5488231
## VD            1    338   338.1   21.3875 3.769e-06 ***
## VE            1     35    35.0    2.2140 0.1367715
## VF            1      3     2.8    0.1767 0.6742026
## VG            1    130   130.3    8.2452 0.0040892 **
## VH            1    217   217.4   13.7498 0.0002093 ***
## VI            1      0     0.1    0.0089 0.9249186
## VK            1      3     2.8    0.1792 0.6720921
## VL            1     18    17.5    1.1095 0.2921964
## VM            1      2     2.2    0.1376 0.7106373
## VN            1     54    54.2    3.4262 0.0641784 .
## VP            1      1     0.6    0.0393 0.8428126
## VQ            1     12    12.0    0.7618 0.3827701
## VR            1     51    50.6    3.2010 0.0736033 .
## VS            1    107   106.9    6.7606 0.0093243 **
## VT            1     13    13.0    0.8242 0.3639730
## VV            1     47    46.8    2.9605 0.0853341 .
## VW            1    262   262.0   16.5703 4.701e-05 ***
## VY            1    993   993.2   62.8254 2.344e-15 ***
## WA            1    155   155.3    9.8224 0.0017258 **
## WC            1     44    44.5    2.8124 0.0935519 .
## WD            1     23    23.5    1.4859 0.2228587
## WE            1     58    58.3    3.6869 0.0548507 .
## WF            1    253   252.9   15.9943 6.370e-05 ***
## WG            1    589   589.4   37.2796 1.037e-09 ***
## WH            1    105   104.6    6.6149 0.0101180 *
## WI            1      2     2.3    0.1456 0.7027393
## WK            1    187   187.4   11.8524 0.0005767 ***
## WL            1    164   163.6   10.3464 0.0012987 **
## WM            1    326   325.7   20.6003 5.683e-06 ***
## WN            1    632   631.7   39.9569 2.636e-10 ***
## WP            1     95    94.9    6.0048 0.0142731 *
## WQ            1     24    24.0    1.5209 0.2174927
## WR            1     25    25.5    1.6107 0.2043996
```

```
## WS           1      0     0.0    0.0000 0.9968910
## WT           1     48    48.5    3.0673 0.0798946 .
## WV           1     33    32.7    2.0689 0.1503425
## WW           1     47    46.8    2.9618 0.0852631 .
## WY           1     39    38.6    2.4433 0.1180359
## YA           1     50    50.0    3.1633 0.0753202 .
## YC           1    181   180.5   11.4205 0.0007274 ***
## YD           1    316   316.1   19.9924 7.806e-06 ***
## YE           1     20    20.3    1.2845 0.2570662
## YF           1      7     7.0    0.4403 0.5069697
## YG           1      4     4.0    0.2534 0.6147157
## YH           1     19    19.1    1.2071 0.2719176
## YI           1    637   637.2   40.3084 2.203e-10 ***
## YK           1      1     0.8    0.0510 0.8214071
## YL           1     97    96.9    6.1300 0.0132967 *
## YM           1    186   185.7   11.7468 0.0006104 ***
## YN           1    131   130.7    8.2648 0.0040451 **
## YP           1      1     0.8    0.0475 0.8274700
## YQ           1     37    36.5    2.3092 0.1286198
## YR           1    714   714.2   45.1757 1.837e-11 ***
## YS           1    220   220.4   13.9389 0.0001892 ***
## YT           1     38    37.5    2.3745 0.1233404
## YV           1    190   190.1   12.0216 0.0005267 ***
## YW           1    171   170.6   10.7922 0.0010206 **
## YY           1    168   167.6   10.5986 0.0011331 **
## pH           1   5953  5952.8  376.5402 < 2.2e-16 ***
## AA:pH        1     33    33.4    2.1135 0.1460210
## AC:pH        1    288   288.3   18.2390 1.955e-05 ***
## AD:pH        1   2934  2933.9  185.5791 < 2.2e-16 ***
## AE:pH        1   5771  5771.2  365.0545 < 2.2e-16 ***
## AF:pH        1      1     1.1    0.0680 0.7943316
## AG:pH        1     60    60.1    3.8021 0.0511981 .
## AH:pH        1     16    15.5    0.9821 0.3216831
## AI:pH        1    203   203.4   12.8631 0.0003357 ***
## AK:pH        1      1     0.7    0.0469 0.8286154
## AL:pH        1    128   127.5    8.0665 0.0045125 **
## AM:pH        1    150   149.9    9.4800 0.0020793 **
## AN:pH        1     10     9.9    0.6259 0.4288875
## AP:pH        1      2     1.5    0.0954 0.7574118
## AQ:pH        1   2486  2485.9  157.2447 < 2.2e-16 ***
## AR:pH        1    161   160.8   10.1710 0.0014283 **
## AS:pH        1     35    35.1    2.2204 0.1362078
## AT:pH        1   3074  3074.0  194.4461 < 2.2e-16 ***
## AV:pH        1     12    12.1    0.7667 0.3812559
## AW:pH        1    273   272.8   17.2549 3.278e-05 ***
## AY:pH        1    394   394.1   24.9267 5.992e-07 ***
## CA:pH        1     80    80.3    5.0782 0.0242369 *
## CC:pH        1     42    42.0    2.6545 0.1032701
## CD:pH        1    332   331.7   20.9793 4.663e-06 ***
## CE:pH        1     20    19.6    1.2387 0.2657288
## CF:pH        1     25    25.3    1.5990 0.2060576
## CG:pH        1    133   132.9    8.4051 0.0037447 **
## CH:pH        1     18    17.8    1.1276 0.2883059
## CI:pH        1    783   782.6   49.5027 2.027e-12 ***
## CK:pH        1    631   631.1   39.9198 2.687e-10 ***
## CL:pH        1      3     3.3    0.2075 0.6487711
## CM:pH        1    935   935.4   59.1682 1.496e-14 ***
## CN:pH        1    231   231.0   14.6089 0.0001326 ***
## CP:pH        1   2096  2095.6  132.5545 < 2.2e-16 ***
## CQ:pH        1     18    18.4    1.1668 0.2800645
```

```
## CR:pH         1    296   295.6   18.6965 1.538e-05 ***
## CS:pH         1    302   302.0   19.1008 1.244e-05 ***
## CT:pH         1   1509  1509.2   95.4606 < 2.2e-16 ***
## CV:pH         1    633   633.1   40.0471 2.517e-10 ***
## CW:pH         1   1061  1060.8   67.1028 2.688e-16 ***
## CY:pH         1     66    65.9    4.1702 0.0411495 *
## DA:pH         1    361   361.3   22.8547 1.756e-06 ***
## DC:pH         1    378   378.0   23.9126 1.014e-06 ***
## DD:pH         1   1818  1817.8  114.9813 < 2.2e-16 ***
## DE:pH         1    106   105.5    6.6757 0.0097788 **
## DF:pH         1    365   364.5   23.0588 1.580e-06 ***
## DG:pH         1    820   819.6   51.8411 6.173e-13 ***
## DH:pH         1     84    84.4    5.3418 0.0208272 *
## DI:pH         1   1147  1147.2   72.5656 < 2.2e-16 ***
## DK:pH         1    282   281.9   17.8300 2.423e-05 ***
## DL:pH         1    295   294.5   18.6290 1.593e-05 ***
## DM:pH         1     74    74.2    4.6912 0.0303266 *
## DN:pH         1   1266  1265.7   80.0611 < 2.2e-16 ***
## DP:pH         1      0     0.5    0.0288 0.8653032
## DQ:pH         1    409   409.4   25.8947 3.629e-07 ***
## DR:pH         1    280   279.8   17.6959 2.600e-05 ***
## DS:pH         1     25    25.2    1.5919 0.2070673
## DT:pH         1     26    25.7    1.6277 0.2020337
## DV:pH         1      1     0.7    0.0458 0.8304871
## DW:pH         1     95    95.4    6.0348 0.0140329 *
## DY:pH         1    155   155.5    9.8338 0.0017151 **
## EA:pH         1    479   479.0   30.2998 3.735e-08 ***
## EC:pH         1   2239  2238.6  141.6037 < 2.2e-16 ***
## ED:pH         1    664   663.8   41.9905 9.327e-11 ***
## EE:pH         1    559   558.7   35.3413 2.800e-09 ***
## EF:pH         1    346   345.6   21.8620 2.944e-06 ***
## EG:pH         1    276   275.7   17.4400 2.974e-05 ***
## EH:pH         1      1     0.9    0.0599 0.8066095
## EI:pH         1    328   327.8   20.7376 5.290e-06 ***
## EK:pH         1    438   437.9   27.6992 1.428e-07 ***
## EL:pH         1    186   186.1   11.7689 0.0006032 ***
## EM:pH         1     85    84.8    5.3660 0.0205398 *
## EN:pH         1     58    58.5    3.6996 0.0544364 .
## EP:pH         1    511   510.5   32.2928 1.339e-08 ***
## EQ:pH         1    501   501.0   31.6900 1.826e-08 ***
## ER:pH         1      0     0.2    0.0110 0.9165341
## ES:pH         1     50    50.3    3.1815 0.0744857 .
## ET:pH         1    690   690.4   43.6695 3.959e-11 ***
## EV:pH         1      0     0.1    0.0064 0.9360508
## EW:pH         1      1     1.3    0.0795 0.7780177
## EY:pH         1      3     2.9    0.1818 0.6698356
## FA:pH         1    415   415.3   26.2680 2.992e-07 ***
## FC:pH         1    253   253.0   16.0037 6.339e-05 ***
## FD:pH         1    142   141.7    8.9658 0.0027532 **
## FE:pH         1     83    83.1    5.2558 0.0218815 *
## FF:pH         1    366   366.2   23.1664 1.494e-06 ***
## FG:pH         1     11    11.4    0.7234 0.3950432
## FH:pH         1      2     1.8    0.1149 0.7346578
## FI:pH         1     90    90.4    5.7175 0.0168032 *
## FK:pH         1     50    50.5    3.1935 0.0739431 .
## FL:pH         1    319   318.9   20.1701 7.114e-06 ***
## FM:pH         1     38    38.1    2.4106 0.1205250
## FN:pH         1     59    59.1    3.7411 0.0531000 .
## FP:pH         1      3     3.3    0.2075 0.6487306
## FQ:pH         1    184   184.5   11.6676 0.0006369 ***
```

```
## FR:pH         1      6     6.5    0.4085 0.5227205
## FS:pH         1    168   167.7   10.6060 0.0011286 **
## FT:pH         1    203   202.8   12.8274 0.0003422 ***
## FV:pH         1     48    48.4    3.0635 0.0800806 .
## FW:pH         1     10     9.7    0.6147 0.4330157
## FY:pH         1     88    87.9    5.5622 0.0183591 *
## GA:pH         1    349   349.5   22.1050 2.594e-06 ***
## GC:pH         1     16    16.2    1.0247 0.3114091
## GD:pH         1     52    52.2    3.3038 0.0691308 .
## GE:pH         1      1     0.7    0.0465 0.8293186
## GF:pH         1     34    34.3    2.1680 0.1409168
## GG:pH         1     25    24.9    1.5756 0.2094134
## GH:pH         1      2     1.6    0.0998 0.7520305
## GI:pH         1     37    37.4    2.3626 0.1242881
## GK:pH         1    151   150.8    9.5363 0.0020165 **
## GL:pH         1    160   160.2   10.1334 0.0014577 **
## GM:pH         1    492   492.3   31.1433 2.419e-08 ***
## GN:pH         1     88    88.4    5.5939 0.0180298 *
## GP:pH         1    227   227.1   14.3657 0.0001508 ***
## GQ:pH         1    463   463.0   29.2881 6.290e-08 ***
## GR:pH         1    260   259.8   16.4310 5.059e-05 ***
## GS:pH         1      3     3.3    0.2076 0.6486319
## GT:pH         1      7     7.4    0.4685 0.4936849
## GV:pH         1     33    32.9    2.0828 0.1489807
## GW:pH         1      8     7.8    0.4960 0.4812688
## GY:pH         1    630   629.8   39.8388 2.800e-10 ***
## HA:pH         1      2     2.1    0.1311 0.7173306
## HC:pH         1     15    14.9    0.9452 0.3309525
## HD:pH         1     50    50.0    3.1643 0.0752763 .
## HE:pH         1     19    19.3    1.2185 0.2696623
## HF:pH         1    247   246.9   15.6204 7.761e-05 ***
## HG:pH         1    138   137.8    8.7187 0.0031523 **
## HH:pH         1     57    57.2    3.6184 0.0571546 .
## HI:pH         1      2     2.1    0.1319 0.7165233
## HK:pH         1     15    15.0    0.9518 0.3292683
## HL:pH         1     41    40.8    2.5776 0.1083942
## HM:pH         1    151   150.9    9.5444 0.0020076 **
## HN:pH         1    515   515.3   32.5980 1.145e-08 ***
## HP:pH         1    108   108.3    6.8514 0.0088618 **
## HQ:pH         1     48    47.7    3.0187 0.0823209 .
## HR:pH         1     48    47.5    3.0077 0.0828804 .
## HS:pH         1    214   213.9   13.5327 0.0002349 ***
## HT:pH         1     16    15.8    0.9965 0.3181756
## HV:pH         1      0     0.1    0.0045 0.9464847
## HW:pH         1     40    40.0    2.5283 0.1118350
## HY:pH         1    214   214.4   13.5638 0.0002310 ***
## IA:pH         1    519   519.4   32.8523 1.005e-08 ***
## IC:pH         1      2     2.4    0.1544 0.6943650
## ID:pH         1    155   155.1    9.8124 0.0017352 **
## IE:pH         1     83    83.3    5.2701 0.0217028 *
## IF:pH         1      1     1.2    0.0776 0.7805902
## IG:pH         1    110   109.9    6.9494 0.0083896 **
## IH:pH         1      9     9.3    0.5857 0.4440869
## II:pH         1    215   215.5   13.6304 0.0002230 ***
## IK:pH         1     47    47.1    2.9793 0.0843495 .
## IL:pH         1     51    51.3    3.2464 0.0715912 .
## IM:pH         1      5     5.4    0.3392 0.5602739
## IN:pH         1     34    33.8    2.1393 0.1435758
## IP:pH         1     11    10.8    0.6839 0.4082555
## IQ:pH         1    573   573.4   36.2723 1.738e-09 ***
```

```
## IR:pH         1    169   169.3   10.7059 0.0010693 **
## IS:pH         1     55    55.0    3.4769 0.0622411 .
## IT:pH         1     12    12.1    0.7655 0.3816186
## IV:pH         1      7     7.0    0.4414 0.5064648
## IW:pH         1    284   283.8   17.9485 2.277e-05 ***
## IY:pH         1      7     6.5    0.4136 0.5201712
## KA:pH         1    592   592.2   37.4618 9.448e-10 ***
## KC:pH         1      7     7.2    0.4554 0.4997648
## KD:pH         1      3     3.2    0.2031 0.6522611
## KE:pH         1     64    64.4    4.0713 0.0436274 *
## KF:pH         1     43    43.3    2.7381 0.0979949 .
## KG:pH         1    212   212.1   13.4182 0.0002497 ***
## KH:pH         1    112   112.0    7.0854 0.0077761 **
## KI:pH         1     82    82.0    5.1864 0.0227715 *
## KK:pH         1     48    48.2    3.0489 0.0808015 .
## KL:pH         1    103   102.6    6.4924 0.0108391 *
## KM:pH         1    223   223.1   14.1107 0.0001727 ***
## KN:pH         1     57    57.1    3.6094 0.0574653 .
## KP:pH         1     52    52.1    3.2955 0.0694817 .
## KQ:pH         1    256   256.2   16.2083 5.690e-05 ***
## KR:pH         1    139   139.0    8.7938 0.0030251 **
## KS:pH         1    102   101.9    6.4479 0.0111136 *
## KT:pH         1    364   363.9   23.0191 1.612e-06 ***
## KV:pH         1      1     1.1    0.0670 0.7957793
## KW:pH         1      2     1.9    0.1211 0.7278337
## KY:pH         1     40    39.7    2.5096 0.1131659
## LA:pH         1     13    13.3    0.8392 0.3596429
## LC:pH         1    218   218.0   13.7869 0.0002052 ***
## LD:pH         1    288   287.7   18.1973 1.998e-05 ***
## LE:pH         1    482   482.2   30.5025 3.365e-08 ***
## LF:pH         1    149   148.7    9.4056 0.0021653 **
## LG:pH         1    109   109.0    6.8974 0.0086370 **
## LH:pH         1     33    33.2    2.0998 0.1473314
## LI:pH         1     46    46.0    2.9127 0.0878940 .
## LK:pH         1   1019  1019.0   64.4562 1.026e-15 ***
## LL:pH         1    273   273.2   17.2833 3.230e-05 ***
## LM:pH         1     51    50.8    3.2140 0.0730237 .
## LN:pH         1      3     2.8    0.1755 0.6753134
## LP:pH         1     43    43.2    2.7295 0.0985243 .
## LQ:pH         1     12    11.5    0.7275 0.3936961
## LR:pH         1    716   716.3   45.3066 1.718e-11 ***
## LS:pH         1      9     9.2    0.5791 0.4466893
## LT:pH         1      4     4.3    0.2702 0.6032246
## LV:pH         1    725   724.5   45.8283 1.317e-11 ***
## LW:pH         1      2     1.6    0.1037 0.7474191
## LY:pH         1     28    28.1    1.7786 0.1823351
## MA:pH         1     20    20.5    1.2955 0.2550483
## MC:pH         1    519   518.9   32.8232 1.020e-08 ***
## MD:pH         1    101   101.2    6.4035 0.0113951 *
## ME:pH         1     82    81.6    5.1618 0.0230962 *
## MF:pH         1     22    21.9    1.3878 0.2387946
## MG:pH         1    196   195.6   12.3738 0.0004361 ***
## MH:pH         1     40    40.4    2.5575 0.1097861
## MI:pH         1    406   406.5   25.7116 3.990e-07 ***
## MK:pH         1    629   629.2   39.7981 2.859e-10 ***
## ML:pH         1     12    12.2    0.7723 0.3795043
## MM:pH         1    396   396.4   25.0728 5.555e-07 ***
## MN:pH         1    144   143.5    9.0779 0.0025895 **
## MP:pH         1     73    73.3    4.6343 0.0313482 *
## MQ:pH         1     96    96.1    6.0783 0.0136912 *
```

```
## MR:pH         1      5     4.9    0.3126 0.5761172
## MS:pH         1     63    62.8    3.9696 0.0463377 *
## MT:pH         1      1     1.4    0.0878 0.7669824
## MV:pH         1     64    63.6    4.0238 0.0448732 *
## MW:pH         1    252   251.7   15.9181 6.632e-05 ***
## MY:pH         1     58    58.4    3.6971 0.0545178 .
## NA.:pH        1    520   520.1   32.9006 9.801e-09 ***
## NC:pH         1     41    41.2    2.6088 0.1062828
## ND:pH         1     26    25.5    1.6142 0.2039110
## NE:pH         1    659   658.9   41.6785 1.094e-10 ***
## NF:pH         1    171   171.0   10.8194 0.0010057 **
## NG:pH         1     12    12.2    0.7717 0.3797080
## NH:pH         1      7     7.3    0.4600 0.4976496
## NI:pH         1    277   276.8   17.5086 2.869e-05 ***
## NK:pH         1      2     1.6    0.0987 0.7533483
## NL:pH         1     35    34.9    2.2081 0.1373026
## NM:pH         1     56    55.5    3.5126 0.0609139 .
## NN:pH         1    137   137.5    8.6959 0.0031919 **
## NP:pH         1      4     3.9    0.2487 0.6179957
## NQ:pH         1     90    90.2    5.7040 0.0169334 *
## NR:pH         1    110   109.7    6.9365 0.0084502 **
## NS:pH         1    225   224.8   14.2206 0.0001629 ***
## NT:pH         1     72    72.4    4.5769 0.0324142 *
## NV:pH         1     68    68.5    4.3313 0.0374275 *
## NW:pH         1     19    19.2    1.2157 0.2702208
## NY:pH         1      2     2.2    0.1379 0.7103811
## PA:pH         1     51    51.5    3.2549 0.0712216 .
## PC:pH         1     59    59.4    3.7545 0.0526772 .
## PD:pH         1     60    59.8    3.7799 0.0518811 .
## PE:pH         1      7     7.0    0.4426 0.5058719
## PF:pH         1      1     1.1    0.0692 0.7924858
## PG:pH         1     42    42.3    2.6749 0.1019540
## PH:pH         1      5     5.0    0.3171 0.5733595
## PI:pH         1      0     0.0    0.0003 0.9867770
## PK:pH         1      4     4.3    0.2733 0.6011248
## PL:pH         1     54    54.5    3.4455 0.0634336 .
## PM:pH         1    263   263.2   16.6514 4.504e-05 ***
## PN:pH         1     52    52.0    3.2891 0.0697537 .
## PP:pH         1    325   325.3   20.5746 5.760e-06 ***
## PQ:pH         1     54    54.5    3.4453 0.0634426 .
## PR:pH         1     38    37.7    2.3865 0.1224025
## PS:pH         1     22    22.3    1.4084 0.2353377
## PT:pH         1      1     0.7    0.0457 0.8307716
## PV:pH         1     15    14.9    0.9397 0.3323592
## PW:pH         1      0     0.0    0.0030 0.9563054
## PY:pH         1    114   114.4    7.2391 0.0071377 **
## QA:pH         1   1565  1565.0   98.9943 < 2.2e-16 ***
## QC:pH         1     67    67.0    4.2371 0.0395580 *
## QD:pH         1     49    49.2    3.1142 0.0776257 .
## QE:pH         1    126   126.5    8.0011 0.0046783 **
## QF:pH         1     14    14.4    0.9100 0.3401266
## QG:pH         1     36    35.5    2.2468 0.1339053
## QH:pH         1    307   307.5   19.4493 1.037e-05 ***
## QI:pH         1    135   135.4    8.5669 0.0034261 **
## QK:pH         1    192   191.8   12.1320 0.0004964 ***
## QL:pH         1     38    37.7    2.3821 0.1227448
## QM:pH         1     40    40.4    2.5555 0.1099199
## QN:pH         1     31    31.5    1.9901 0.1583453
## QP:pH         1     35    34.8    2.2000 0.1380248
## QQ:pH         1      0     0.1    0.0054 0.9411963
```

```
## QR:pH         1     29    29.1    1.8422 0.1747026
## QS:pH         1    217   217.4   13.7507 0.0002092 ***
## QT:pH         1    104   103.8    6.5645 0.0104083 *
## QV:pH         1    139   138.7    8.7710 0.0030632 **
## QW:pH         1     58    58.0    3.6673 0.0555003 .
## QY:pH         1    319   319.0   20.1758 7.093e-06 ***
## RA:pH         1    122   121.6    7.6919 0.0055506 **
## RC:pH         1      1     1.1    0.0681 0.7941428
## RD:pH         1     19    18.6    1.1791 0.2775444
## RE:pH         1    306   306.2   19.3662 1.083e-05 ***
## RF:pH         1     46    45.9    2.9039 0.0883785 .
## RG:pH         1    163   162.5   10.2819 0.0013450 **
## RH:pH         1     36    35.6    2.2517 0.1334767
## RI:pH         1     69    68.8    4.3501 0.0370157 *
## RK:pH         1     15    14.9    0.9436 0.3313573
## RL:pH         1      0     0.1    0.0047 0.9455467
## RM:pH         1     11    11.3    0.7127 0.3985633
## RN:pH         1      5     4.9    0.3110 0.5770689
## RP:pH         1    211   211.0   13.3469 0.0002593 ***
## RQ:pH         1     23    22.9    1.4515 0.2282966
## RR:pH         1    211   210.8   13.3345 0.0002611 ***
## RS:pH         1    130   130.3    8.2418 0.0040969 **
## RT:pH         1     96    95.6    6.0489 0.0139211 *
## RV:pH         1     38    37.8    2.3930 0.1218936
## RW:pH         1     48    48.5    3.0660 0.0799565 .
## RY:pH         1    170   170.5   10.7836 0.0010253 **
## SA:pH         1    109   109.4    6.9199 0.0085290 **
## SC:pH         1     69    69.5    4.3947 0.0360603 *
## SD:pH         1      3     3.3    0.2086 0.6478743
## SE:pH         1    187   187.5   11.8586 0.0005748 ***
## SF:pH         1     60    59.8    3.7814 0.0518368 .
## SG:pH         1    177   177.0   11.1974 0.0008202 ***
## SH:pH         1    151   151.4    9.5771 0.0019722 **
## SI:pH         1    301   301.3   19.0555 1.274e-05 ***
## SK:pH         1      1     1.4    0.0858 0.7695507
## SL:pH         1      2     1.9    0.1223 0.7265312
## SM:pH         1    379   378.7   23.9556 9.914e-07 ***
## SN:pH         1    425   424.8   26.8700 2.192e-07 ***
## SP:pH         1     41    40.8    2.5819 0.1081060
## SQ:pH         1      4     4.0    0.2547 0.6137736
## SR:pH         1    150   150.0    9.4887 0.0020694 **
## SS:pH         1     95    95.4    6.0370 0.0140150 *
## ST:pH         1     26    26.1    1.6496 0.1990270
## SV:pH         1      1     1.3    0.0825 0.7739666
## SW:pH         1      7     6.6    0.4145 0.5197233
## SY:pH         1      0     0.1    0.0094 0.9229578
## TA:pH         1     20    20.0    1.2630 0.2611010
## TC:pH         1     55    55.4    3.5074 0.0611049 .
## TD:pH         1     31    30.6    1.9357 0.1641428
## TE:pH         1      1     1.3    0.0793 0.7782802
## TF:pH         1    132   132.2    8.3596 0.0038396 **
## TG:pH         1    145   145.4    9.1996 0.0024230 **
## TH:pH         1      1     1.0    0.0602 0.8061441
## TI:pH         1      0     0.1    0.0073 0.9320435
## TK:pH         1      4     4.4    0.2795 0.5970223
## TL:pH         1    400   400.3   25.3198 4.888e-07 ***
## TM:pH         1      0     0.0    0.0030 0.9565894
## TN:pH         1     22    22.3    1.4131 0.2345482
## TP:pH         1      0     0.5    0.0285 0.8659747
## TQ:pH         1      1     0.7    0.0431 0.8354566
```

```
## TR:pH         1     95    95.5    6.0377 0.0140100 *
## TS:pH         1     17    16.5    1.0450 0.3066803
## TT:pH         1     68    67.8    4.2856 0.0384462 *
## TV:pH         1      4     3.8    0.2388 0.6250890
## TW:pH         1     34    33.7    2.1341 0.1440653
## TY:pH         1      0     0.5    0.0316 0.8588902
## VA:pH         1      7     6.8    0.4294 0.5123071
## VC:pH         1     57    57.2    3.6169 0.0572064 .
## VD:pH         1     40    39.8    2.5186 0.1125246
## VE:pH         1      3     2.6    0.1669 0.6829073
## VF:pH         1      4     4.1    0.2582 0.6113792
## VG:pH         1     11    11.2    0.7087 0.3998865
## VH:pH         1    188   188.4   11.9195 0.0005563 ***
## VI:pH         1     20    19.8    1.2524 0.2631030
## VK:pH         1    238   238.0   15.0575 0.0001045 ***
## VL:pH         1    402   402.2   25.4439 4.583e-07 ***
## VM:pH         1    151   150.9    9.5469 0.0020049 **
## VN:pH         1    136   135.7    8.5861 0.0033901 **
## VP:pH         1      0     0.0    0.0030 0.9564121
## VQ:pH         1      7     7.2    0.4539 0.5004890
## VR:pH         1     12    12.2    0.7700 0.3802103
## VS:pH         1     56    56.5    3.5708 0.0588128 .
## VT:pH         1      2     1.5    0.0980 0.7542374
## VV:pH         1      9     8.8    0.5589 0.4547051
## VW:pH         1      5     5.3    0.3348 0.5628621
## VY:pH         1     31    30.6    1.9361 0.1641075
## WA:pH         1     13    13.4    0.8460 0.3577013
## WC:pH         1      9     9.5    0.6005 0.4383878
## WD:pH         1    175   175.3   11.0862 0.0008709 ***
## WE:pH         1     17    17.5    1.1046 0.2932739
## WF:pH         1     13    12.9    0.8148 0.3667020
## WG:pH         1      1     1.2    0.0748 0.7844281
## WH:pH         1     48    47.6    3.0120 0.0826604 .
## WI:pH         1     66    66.3    4.1960 0.0405283 *
## WK:pH         1      4     4.2    0.2646 0.6069491
## WL:pH         1      5     5.2    0.3280 0.5668226
## WM:pH         1    148   148.1    9.3649 0.0022139 **
## WN:pH         1      0     0.3    0.0166 0.8975302
## WP:pH         1    132   132.5    8.3791 0.0037986 **
## WQ:pH         1     99    99.1    6.2711 0.0122781 *
## WR:pH         1    436   435.9   27.5730 1.524e-07 ***
## WS:pH         1     34    34.0    2.1494 0.1426382
## WT:pH         1     54    53.6    3.3934 0.0654688 .
## WV:pH         1    122   121.6    7.6925 0.0055488 **
## WW:pH         1     51    51.3    3.2431 0.0717359 .
## WY:pH         1     22    21.9    1.3863 0.2390347
## YA:pH         1      8     8.4    0.5284 0.4672793
## YC:pH         1    137   136.6    8.6409 0.0032896 **
## YD:pH         1     66    65.8    4.1647 0.0412835 *
## YE:pH         1     21    21.3    1.3494 0.2454012
## YF:pH         1    219   218.6   13.8259 0.0002009 ***
## YG:pH         1      1     1.2    0.0748 0.7844974
## YH:pH         1     44    44.2    2.7929 0.0946949 .
## YI:pH         1     61    61.5    3.8874 0.0486600 *
## YK:pH         1    131   130.8    8.2731 0.0040268 **
## YL:pH         1      3     3.2    0.2012 0.6537733
## YM:pH         1    823   822.9   52.0496 5.552e-13 ***
## YN:pH         1      1     1.1    0.0681 0.7940932
## YP:pH         1     17    17.1    1.0790 0.2989239
## YQ:pH         1      2     2.5    0.1579 0.6910994
```

```
## YR:pH         1      1     0.8    0.0512 0.8209751
## YS:pH         1     33    33.2    2.0979 0.1475160
## YT:pH         1    113   112.7    7.1278 0.0075944 **
## YV:pH         1      1     0.6    0.0375 0.8465135
## YW:pH         1     18    17.6    1.1120 0.2916522
## YY:pH         1      0     0.2    0.0112 0.9156354
## Residuals 27449 433946    15.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix2: All code for this report

```
knitr::opts_chunk$set(echo = TRUE)
library(Matrix)
library(psych)
library(hash)
library(MASS)

library(stringr)
library(optimx)
library(leaps)
training_data <- read.csv(file = 'train.csv')
keeps <- c("protein_sequence", "pH","tm")
training_data <- training_data[keeps]

update_training_data <- read.csv(file = 'train_updates_20220929.csv')
for (row in 1:nrow(update_training_data)) {
  if(update_training_data[row, "protein_sequence"] != "") {
    training_data[row, "protein_sequence"] <- update_training_data[row, "protein_sequence"]
    training_data[row, "pH"] <- update_training_data[row, "pH"]
    training_data[row, "tm"] <- update_training_data[row, "tm"]
  }
}

#sapply(training_data,class)

#print(nrow(training_data))
N <- nrow(training_data)
testing_data <- read.csv(file = 'test.csv')
#print(nrow(testing_data))
M <- nrow(testing_data)
training_data$pH[is.na(training_data$pH)] <- 7

print_hist <- function(x, title, scale) {

  x2 <- seq(min(x), max(x), length = 40)

  fun <- dnorm(x2, mean = mean(x), sd = sd(x))

  hist(x, prob = TRUE, ylim = c(0, scale), col = "white", main=title)
  lines(density(x), col = 4, lwd = 2)
}

par(mfrow = c(2, 2))

print("tm quartiles:")
quantile(training_data$tm)

print("pH quartiles:")
```

```r
quantile(training_data$pH)

print("Protein Sequence Length quartiles:")
quantile(nchar(training_data$protein_sequence))

print_hist(training_data$tm, 'Histogram of Thermostability',0.06)

print_hist(training_data$pH, 'Histogram of pH Level',0.3)

print_hist(nchar(training_data$protein_sequence), 'Histogram of Protein Length',0.0025)



# lambda <- bc$x[which.max(bc$y)]
# print(lambda)
#
# K_2 <- prod(Y)^(1/length(Y))
# K_1 <- 1/(K_2^(lambda - 1))
# print(K_1)
#
# Y_transform <- K_1/lambda * (Y^lambda - 1)
# inverse_transform <- function(z) {
#   return((lambda*z / K_1 + 1)^(1/lambda))
# }
knitr::include_graphics("nonlinear.png")

phrase_2_col_dictionary <- hash()

states <- c('A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y')
col_names <- rep("", 402)
#col_names <- rep("", 8002)
col_names[1] <- "pH"

col = 2
for (i in states) {
  for ( j in states ) {
    #for (k in states) {
      #phrase_2_col_dictionary[[paste(i,j,k,sep='')]] = col
      #col_names[col] <- paste(i,j,k,sep='')
      phrase_2_col_dictionary[[paste(i,j,sep='')]] = col
      col_names[col] <- paste(i,j,sep='')
      col <- col + 1
    #}
  }
}

col_names[402] <- "tm"
#
# training_data_wide <- data.frame(matrix(ncol = 402, nrow = 0))
#
# # col_names[8002] <- "tm"
# #
# # training_data_wide <- data.frame(matrix(ncol = 8002, nrow = 0))
# colnames(training_data_wide) <-col_names
#
# for (row in 1:N) {
#   curr_sequence <- training_data[row, "protein_sequence"]
#   #training_data_wide[row,] <- rep(0, 8002)
#   training_data_wide[row,] <- rep(0, 402)
```

```
#   training_data_wide[row,"pH"] <- training_data[row, "pH"]
#   training_data_wide[row,"tm"] <- training_data[row, "tm"]
#
#   #for (k in 1:(nchar(curr_sequence)-2)) {
#   #  curr_substr <- substr(curr_sequence, k, k+2)
#   for (k in 1:(nchar(curr_sequence)-1)) {
#     curr_substr <- substr(curr_sequence, k, k+1)
#
#     training_data_wide[row,col_names[phrase_2_col_dictionary[[curr_substr]]]] <- training_data_wide[row,col_
#
#   }
# }
#
# write.csv(training_data_wide, "training_data_wide.csv", row.names=FALSE)

training_data_wide <- read.csv(file = 'training_data_wide_ones.csv')
#training_data_wide = subset(training_data_wide, select = -c(X) )
#names(training_data_wide)[names(training_data_wide) == 'NA.'] <- 'NA'

# reduced_df <- training_data_wide[ , -which(names(training_data_wide) %in% c("pH","tm"))]
# reduced_df[-1] <- as.numeric(reduced_df[-1] != 0)
# reduced_df$pH <- training_data_wide$pH
# reduced_df$tm <- training_data_wide$tm
#
# training_data_wide <- reduced_df
#
# write.csv(training_data_wide, "training_data_wide_ones.csv", row.names=FALSE)

shift <- 0.01 - min(training_data_wide$tm)
bc <- boxcox(tm + shift ~ ., data=training_data_wide)

lambda <- bc$x[which.max(bc$y)]
print('lambda')
print(lambda)

K_2 <- 1
K_1 <- 1

Y_transform <- function(z) {
  return(K_1/lambda * ((z + shift)^lambda - 1))
}

inverse_transform <- function(z) {
  return((lambda*(z - shift) / K_1 + 1)^(1/lambda))
}

#write.csv(training_data_wide, "training_data_wide_ones.csv", row.names=FALSE)

#training_data_wide$sequence_lengths <- nchar(training_data$protein_sequence)

#sapply(training_data_wide,class)
set.seed(402)
ind <- sample(1:N, 0.9*N, replace=FALSE)

train <- training_data_wide[ind, ] #training set
valid <- training_data_wide[-ind, ] #validation/test set

train1 <- lm(Y_transform(tm) ~ . + pH:., data = train)
valid1 <- lm(Y_transform(tm) ~ . + pH:., data = valid)
```

```r
model_coefs <- sort(train1$coefficients)

print('Maximum Coefficients')
print(model_coefs[length(model_coefs) - (4:0)])

print('Minimum Coefficients')
print(model_coefs[1:5])

model_coefs <- sort(abs(train1$coefficients))

print('Least Impactful Coefficients')
print(model_coefs[1:5])

par(mfrow = c(2, 2))


plot(train1,which=1) ##residuals vs. fitted values

# train_betas <- coef(summary(train1))[,1]
# valid_beta <- coef(summary(valid1))[,1]
# deltas <- abs(train_betas - valid_beta)

plot(train1,which=2) ##residuals Q-Q plot

# mod_sum <- cbind(coef(summary(train1))[,1], coef(summary(valid1))[,1],
#   coef(summary(train1))[,2], coef(summary(valid1))[,2],deltas)
# colnames(mod_sum) <- c("Train Est","Valid Est","Train s.e.","Valid s.e.","Normed Delta in Coefficients")

boxplot(train1$residuals) ## residuals boxplot




sse_t <- sum(train1$residuals^2)
sse_v <- sum(valid1$residuals^2)
Radj_t <- summary(train1)$adj.r.squared
Radj_v <- summary(valid1)$adj.r.squared
train_sum <- c(sse_t,Radj_t)
valid_sum <- c(sse_v,Radj_v)
criteria <- rbind(train_sum,valid_sum)
colnames(criteria) <- c("SSE","R2_adj")
print(criteria)


#Get MSPE_v from new data
#newdata <- valid[, -1]
y.hat <- predict(train1, valid)
MSPE <- mean((Y_transform(valid$tm) - y.hat)^2)

print('MSPE')
print(MSPE)

print('SSE / N')
print(sse_t/N)
#
#
hist(y.hat, prob = TRUE, ylim = c(0, 0.16), xlab="mean tm contribution", main="Predicted versus True Distribut
lines(density(y.hat), col = 4, lwd = 2)
hist(Y_transform(valid$tm), prob = TRUE, add=TRUE)
```

```r
lines(density(Y_transform(valid$tm)), col = 2, lwd = 2)


summary(train1)
anova(train1)

# testing_data_wide <- data.frame(matrix(ncol = 401, nrow = 0))
# colnames(testing_data_wide) <-col_names[-length(col_names)]
#
# for (row in 1:M) {
#   curr_sequence <- testing_data[row, "protein_sequence"]
#   testing_data_wide[row,] <- rep(0, 401)
#   testing_data_wide[row,"pH"] <- testing_data[row, "pH"]
#
#   for (k in 1:(nchar(curr_sequence)-1)) {
#     curr_substr <- substr(curr_sequence, k, k+1)
#
#     testing_data_wide[row,col_names[phrase_2_col_dictionary[[curr_substr]]]] <- 1
#   }
# }
#
#  write.csv(testing_data_wide, "testing_data_wide.csv", row.names=FALSE)

testing_data_wide <- read.csv("testing_data_wide.csv")

predictions <- predict(train1, testing_data_wide)
pred_Y <- inverse_transform(predictions)
submission_df <- data.frame(matrix(ncol = 2, nrow = 0))
colnames(submission_df) <-c("seq_id", "tm")

for (row in 1:M) {
    submission_df[nrow(submission_df) + 1,] <- c(testing_data[row, "seq_id"], pred_Y[row])
}

write.csv(submission_df,"submission.csv", row.names = FALSE)
```

```python
# Nonlinear Model Python Code

import pandas as pd
from collections import defaultdict
from biopandas.pdb import PandasPdb
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import gc
import warnings
import datetime as dt
import math
from random import sample

np.random.seed(0)
warnings.simplefilter("ignore")
import statistics as stats

from sklearn.model_selection import train_test_split

from statistics import median
from collections import Counter

from sklearn.ensemble import RandomForestRegressor
```

```
from sklearn import metrics

from sklearn.tree import export_graphviz
from sklearn.tree import plot_tree
import graphviz

# # Reading Data Files

train = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
sub = pd.read_csv("sample_submission.csv")

display(train.head())
display(test.head())
display(sub.head())

# update to training data
update = pd.read_csv("train_updates_20220929.csv")
for index, row in update.iterrows():
    if not pd.isnull(row['protein_sequence']):
        train.loc[index,'seq_id'] = update.loc[index,'seq_id']
        train.loc[index,'protein_sequence'] = update.loc[index,'protein_sequence']
        train.loc[index,'pH'] = update.loc[index,'pH']
        train.loc[index,'data_source'] = update.loc[index,'data_source']
        train.loc[index,'tm'] = update.loc[index,'tm']

# data imputation
for index, row in train.iterrows():
    if pd.isnull(row['pH']):
        train.loc[index,'pH'] = 7

protein_sequence_len = []

for sequence in train['protein_sequence'].to_list():
    protein_sequence_len.append(len(sequence))

print("train shape:", train.shape)
print("test shape:", test.shape)
print("sub shape:", sub.shape)

print("train nan value sum:", train.isna().sum().sum())
print("test nan value sum:", test.isna().sum().sum())

train.isna().sum()

Xy = train.to_numpy()

X_train = Xy[:,:-1]
y_train = Xy[:,-1]


# # Analyze 3-gram distributions

n_gram = []
seq_id = []
ph_level = []
origin_freq = []
expected_tm = []
```

```
prevalence = defaultdict(int)
typical_ph_levels = defaultdict(list)


for curr_val, my_tm in zip(X_train, y_train):
    my_id = curr_val[0]
    my_sequence = curr_val[1]
    my_ph = curr_val[2]

    for k in range(len(my_sequence) - 2):
        my_substring = my_sequence[k:k+3]
        n_gram.append(my_substring)
        seq_id.append(my_id)
        ph_level.append(my_ph)
        origin_freq.append(my_sequence.count(my_substring))
        prevalence[my_substring] += 1
        typical_ph_levels[my_substring].append(my_ph)
        expected_tm.append(my_tm)

states = ['A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y']

phrase_2_col_dictionary = {}

substring_length = 3
vector_length = len(states)**substring_length + 2

col_names = []
col_names.append("pH")

col = 1
for i in states:
    for j in states:
        for k in states:
            curr_phrase = "".join([i,j,k])
            phrase_2_col_dictionary[curr_phrase] = col
            col_names.append(curr_phrase)
            col = col + 1

col_names.append("tm")


training_data_wide = pd.DataFrame(columns = col_names)

count = 0
for curr_val, my_tm in zip(X_train, y_train):
    my_id = curr_val[0]
    my_sequence = curr_val[1]
    my_ph = curr_val[2]

    new_vec = [0 for _ in range(vector_length)]
    new_vec[0] = my_ph
    new_vec[vector_length-1] = my_tm

    for k in range(len(my_sequence) - substring_length + 1):
        my_substring = my_sequence[k:k+substring_length]
        new_vec[phrase_2_col_dictionary[my_substring]] += 1

    res = {col_names[i]: new_vec[i] for i in range(len(new_vec))}
    training_data_wide = training_data_wide.append(res, ignore_index = True)
```

```
    count += 1

    if count % 100 == 0:
        print(count)

training_data_wide.to_csv('3_gram_factorization.csv')


sequence_vals = defaultdict(list)
average_ph_levels = defaultdict(float)

for phrase, tm in zip(n_gram, expected_tm):
    sequence_vals[phrase].append(tm)

for key, val in typical_ph_levels.items():
    average_ph_levels[key] = stats.mean(val)

sequence_mean = defaultdict(float)
sequence_median = defaultdict(float)
sequence_std = defaultdict(float)
sequence_max = defaultdict(float)
sequence_min = defaultdict(float)
sequence_statmax = defaultdict(float)
sequence_statmin = defaultdict(float)
median_ph = defaultdict(float)
sequence_prevalance = defaultdict(int)

for phrase, vals in sequence_vals.items():
    sequence_prevalance[phrase] = len(vals)
    sequence_mean[phrase] = stats.mean(vals)
    sequence_median[phrase] = stats.median(vals)
    sequence_std[phrase] = stats.stdev(vals)
    sequence_max[phrase] = max(vals)
    sequence_min[phrase] = min(vals)
    sequence_statmax[phrase] = sequence_mean[phrase] + sequence_std[phrase]
    sequence_statmin[phrase] = sequence_mean[phrase] - sequence_std[phrase]
    median_ph[phrase] = stats.median(typical_ph_levels[phrase])


df = pd.DataFrame([(phrase, sequence_prevalance[phrase], sequence_std[phrase], median_ph[phrase], sequence_mea
                    columns=['subsequence', 'prev', 'std','median_ph','mean_tm','max_tm','min_tm','median_tm','

df.to_csv('phrase_stats.csv')


# # Train - Test Split

X_train, X_val, y_train, y_val = train_test_split(Xy[:,:-1], Xy[:,-1], test_size=0.30, random_state=42)

# # Model Definition

def sigmoid(z):
    if z > 10:
        return 0
    return 1/(1 + math.exp(z))

def predictive_value(alpha,s_i, curr_ph):
    if sequence_prevalance[s_i] == 0:
        return 0
```

```
    return sigmoid(alpha*(sequence_prevalance[s_i])*(1 + sequence_std[s_i])*math.exp(abs(curr_ph - median_ph[s

def expected_value(beta,s_i):
    return_val = beta[0]
    return_val += beta[1]*sequence_mean[s_i]
    return_val += beta[2]*sequence_max[s_i]
    return_val += beta[3]*sequence_min[s_i]
    return_val += beta[4]*sequence_median[s_i]
    return_val += beta[5]*sequence_statmin[s_i]
    return_val += beta[6]*sequence_statmax[s_i]

    return return_val


def predictor(param_vec, curr_sequence, curr_ph):
    num_val = 0
    den_val = 0

    for k in range(len(curr_sequence) - 2):
        s_k = my_sequence[k:k+3]
        pred_val = predictive_value(param_vec[0:4], s_k, curr_ph)
        exp_val = expected_value(param_vec[4:12],s_k)
        num_val += pred_val*exp_val
        den_val += pred_val

        num_val += exp_val
        den_val += pred_val

    if den_val == 0:
        return 0
    else:
        return num_val / len(curr_sequence)

predictor([1,1,1,1,1,1,1,1,1,1,1], X_train[0][1], X_train[0][2])

def Loss(param_vec):

    print('Call to Loss')

    batch_size = 1000

    sampled_list = sample(range(len(X_train)), batch_size)

    my_sum = 0
    count = 0

    for k in sampled_list:
        curr_ph = X_train[k][2]
        curr_sequence = X_train[k][1]
        Y_pred = predictor(param_vec, curr_sequence, curr_ph)
        Y_true = y_train[k]
        my_sum += (Y_pred - Y_true)**2
        count += 1

    return_val = my_sum / count
    print('\tLoss =', return_val)

    return return_val
```

```
Loss([1,1,1,1,1,1,1,1,1,1,1])


# # Scipy Loss Optimization
from scipy.optimize import minimize
x0 = [1,1,1,1,1,1,1,1,1,1,1]
res = minimize(Loss, x0, method='nelder-mead', options={'xatol': 1e-8, 'disp': True,'maxiter':100})

print(res.x)
x0 = res.x

param_vec = res.x


# # Validation Set MSE

my_sum = 0
count = 0
y_val_pred = []

for k in range(len(X_val)):
    curr_ph = X_val[k][2]
    curr_sequence = X_val[k][1]
    Y_pred = predictor(param_vec, curr_sequence, curr_ph)
    y_val_pred.append(Y_pred)
    Y_true = y_val[k]
    my_sum += (Y_pred - Y_true)**2
    count += 1

print(my_sum / count)

bins = np.linspace(-10, 140, 100)
from matplotlib import pyplot

pyplot.hist(y_val, bins, alpha=0.5, label='x')
pyplot.hist(y_val_pred, bins, alpha=0.5, label='y')
pyplot.legend(loc='upper right')
pyplot.show()


# # Submission

submission = defaultdict(float)
for my_id, my_sequence, my_ph in zip(test['seq_id'].tolist(),test['protein_sequence'].tolist(),test['pH'].toli

    submission[my_id] = predictor(param_vec, my_sequence, my_ph)

submission_df = pd.DataFrame(
    {"seq_id": submission.keys(), "tm": submission.values()}
)
submission_df

submission_df.to_csv('submission.csv', index=False)
```

References

EDA Novozymes Enzyme Stability - Eduardo Reyes and Ifeanyi Ezenwa

https://www.kaggle.com/code/seyered/eda-novozymes-enzyme-stability

Datasource

Novozymes Enzyme Stability Prediction Competition Website:

`https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction/overview`