Expmeasure v1.2 User Manual

Statistical Toolkits for Part Characterization

Wu Hao-Ran (zjuhaoranwu@gmail.com)

16 September 2021

Contents

1 Quick-start example

1.1 Test Data

Two datasets are provided in the website https://2021.igem.org/Team:ZJU-China/Software, namely `Prcn_metal.csv` and `sPrcn.csv`. For more details about other datasets, see Section 6.2 and 6.3.

The `Prcn_metal.csv` describes the relationship between the florescence activity of Part *Prcn* (BBa_K540001, see http://parts.igem.org/Part:BBa_K540001) and two factors: types of metal ions and the concentration of ions. The first three rows are shown in table 1.1.1.

Table 1.1.1. First three rows of the test data: `Prcn_metal.csv`

Ion	con	flu	OD600	measure	value
Zn	0	4367	0.401	1	10890.274
Zn	0	10388	0.389	1	26704.37
Zn	0	4609	0.178	1	25893.258

* "Ion": types of metal ion; "con": concentration of metal ion; "flu": florescence; "OD600": optical density of a bacteria medium measured at a wavelength of 600 nm (reflects the number of bacterial cells); "measure": replicates; "value": florescence/OD600 (a measure of the florescence activity)

The `sPrcn.csv` describes the relationship between the florescence activity of Part *sPrcn* (see https://2021.igem.org/Team:ZJU-China/Improve) and various factors: the concentration of Co^{2+}, incubation time, and the concentration of IPTG. The first three rows are shown in table 1.1.2.

Table 1.1.2. First three rows of the test data: `sPrcn.csv`

Co	time	IPTG	OD	GFP	value
0	0	0	0.508	14091	27738.189
0	0	0	0.537	15863	29540.037
0	0	0	0.553	15739	28461.121

1.2 Install

A github version (https://github.com/IGEM-ZJU-2021/expmeasure) is available. You can download and install the R package using the following code in R/Rstudio:

```
install.packages("devtools") # only required if you do not have this package
library(devtools)
install_github("IGEM-ZJU-2021/expmeasure")
```

If you fail to install `Expmeasure`, there is another way to use it (See Appendix1)

1.3 Run `expmeasure()`

To get started with this software, only two commands are required since `Expmeasure` is a user-friendly software that provides GUI interface to all users.

```
library(expmeasure)
expmeasure()
```

You will see the HOME page of our software in the browser that is launched by `Expmeasure` (Fig. 1.3.1).

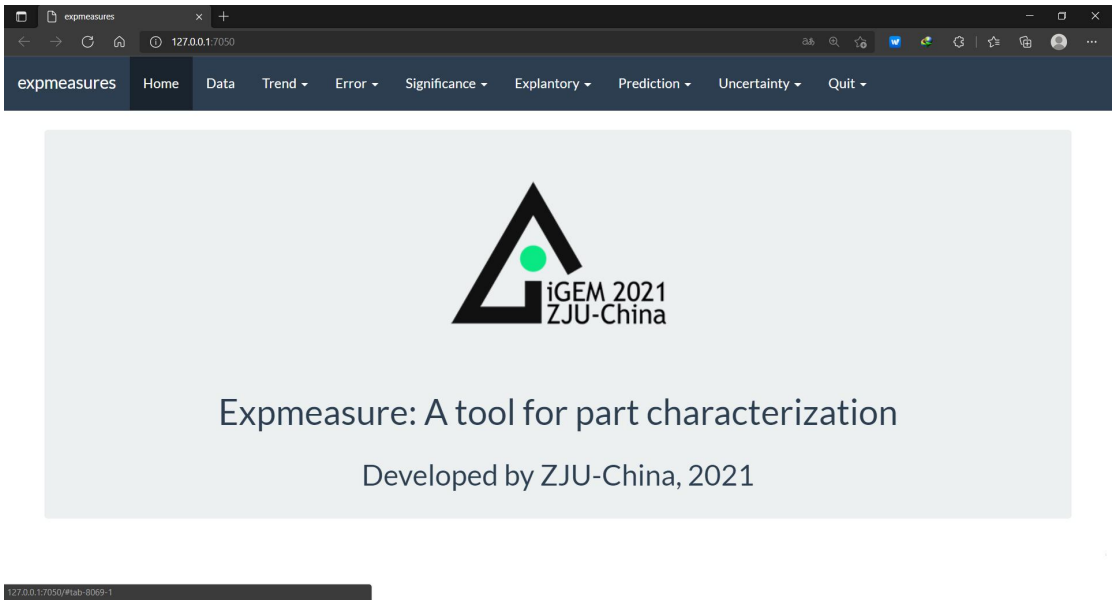

Figure 1.3.1 The HOME page of `Expmeasure`

In the top navigation bar, you can see several icons that allows you to upload data or perform statistical analyses.

Click the "Data" icon at the top of your screen, then it will turn to DATA page (Figure 1.3.2). You will see side bar panel at the left which allows you to send command to the server, and the output panel at the right. Click "Browse..." icon and then you will be required to upload a data file. In this example, I upload `Prcn_metal.csv` data file. After uploading your data, the first six rows of your dataset will represent at the right panel. An error message on the right panel indicates that the format of your data file is not valid. See Chapter 4 for detailed information about data format or try to make you data looks like `Prcn_metal.csv`.

After successfully uploading your data, choose a response variable in the select picker named by "Select response variable". Here I choose "value" which describes the florescence activity of Part *Prcn*. You are required to choose candidate explanatory variables in the next picker called "Select explanatory variables".

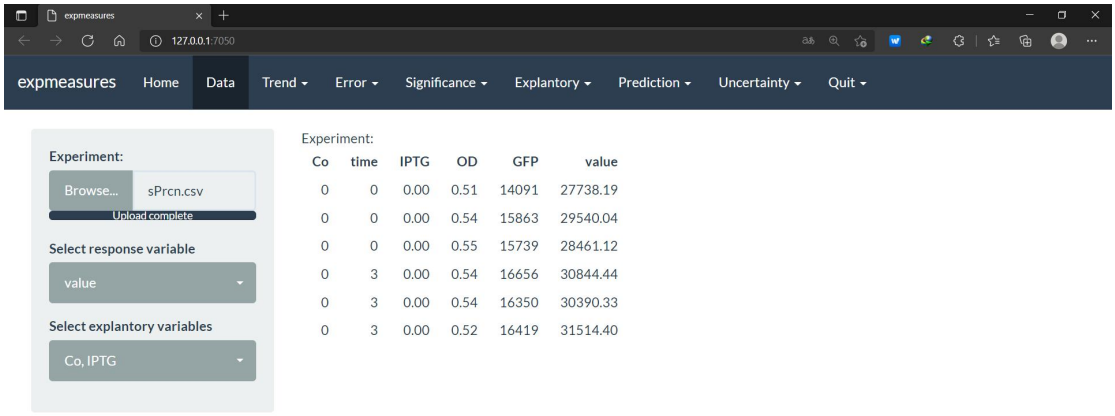

Figure 1.3.2. The DATA page of `Expmeasure`

The next step is to perform all statistical tests available in this software. As an example, turn to "Trend > VioPlot" page. This page helps you to generate a violin plot based on an explanatory variable and then perform t-test across various levels. In the "Select explanatory variable" picker at "VioPlot" page, only "ion" variable makes sense. This is because violin plot is only designed for catogoraical variables. Click "APPLY!" icon and you will see an interactive plot at the right panel (Fig. 1.3.3).

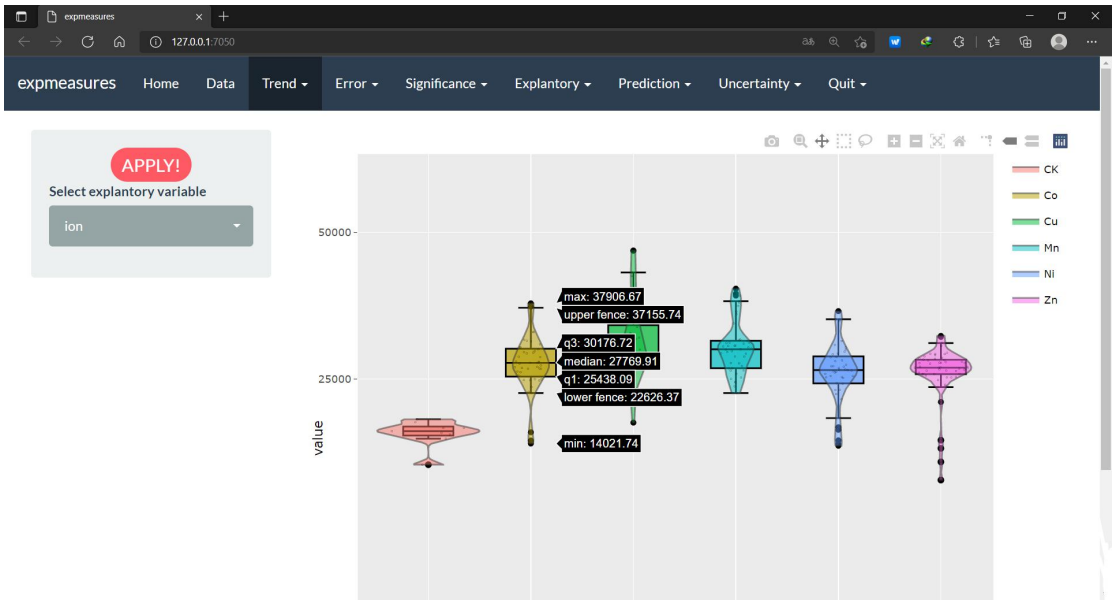

Figure 1.3.3. Violin plot generated by `Expmeasure`

You can benefit from a couple of useful tools (e.g. download picture, zoom in, and zoom out) in the top-right of the picture (Fig. 1.3.3).

1.4 The structure of this manual

The rest of this manual provides detailed information about performing statistical analyses using `Expmeasure` in a sensible way. You may be happy with excellent output in `Expmeasure` without reading the following chapters. But if you come across puzzling error messages, or not sure about the output, please be patient and read corresponding material in the rest of this manual.

Chapter 2 introduces statistical theory for each module. If you are concerned about how the results are generated, check some details in the chapter. Chapter 3 introduces why I design `Expmeasure` and how it can be used for future iGEM teams. A quick user may not think these two chapters are important for them.

Chapter 4 describes the data input in `Expmeasure`. If you have gotten an error message in the DATA page, I strongly recommended you to read this chapter carefully. Chapter 5 gives a briefly overview of each module, including the input information, output, and tips to avoid wrong statistical analyses. It is often useful when you get an error message when performing analyses.

Chapter 6 provides a suggested workflow to use `Expmeasure`. You may be happy at excellent graph generated by `Expmeasure` without going through this chapter. However, to make correct conclusions and useful suggestions for further experiments, sensible use of statistical methods is important. Consequently, it is also not suggested to ignore this chapter unless you are extremely satisfied with the results.

2 Statistical Basics

2.1 Concepts of statistics

In wet lab, we often make conclusions based on an experiment. We detect high intensity of fluorescence activity and thus being confident that our engineered bacteria can express *GFP*. We find a great difference of vital rate after adding drugs and then make a conclusion that only small dose of drug can activate the kill switch in the cell. We see a band when carrying out western blot and then believe that the engineered bacteria expressed the targeted protein. Unfortunately, this might be irreproducible it is a false-positive result. Inaccurate conclusions can be made if we do not realize random errors in experiments.

What if we repeat the same experiment for 10 times and find that nine of them gives us the results that we expected? In this occasion, we are confident that the result is reliable.

Statistics is an area of applied mathematics that helps us to make effective inferences and reasonable conclusions on population based on our dataset. **Population** is a collection of all individuals in an object that we are interested in. For example, the editor of a journal raises questions about your conclusions in your article and suggest you to repeat your experiments. But how many experiments? Obviously, you do not want to repeat your experiments for 1000+ times. Here the editor is interested in results from a collection of experiments (population). But it is impossible for you to survey all individuals in the population. Instead of that, you perform your wet experiments for several times, and that is called **sample**. You collect data from numerous samples and statistical methods help you to make inferences on the properties of population by analyzing data collected from the samples.

To make sensible inferences, characterizing population properties using statistics is important. **Statistics** (not the same concept as the previous one) is a quantity that describes characteristics of a sample or a population. **Models** are then used to simulate our hypothesis. If the value of statistics calculated based on real data do not show a big difference from the values of statistics calculated based on simulated data, we conclude that the model may reflect the true mechanisms in the dataset. If not, the model (or corresponding hypothesis) is rejected.

Many conclusions describe the relationship between two variables. The **response variable** is an indicator that can be influenced by other factors. The **explanatory variables** usually have an impact on the response variable. A variable can be roughly divided into two categories. **Categorical variable** is expressed by different groups, while **continuous variable** is expressed in any real number. In some cases, categorical variable can be further divides into two types. But in `Expmeasure`, I am not going to introduce further information about that.

2.2 Generalized linear model (GLMs)

Generalized linear model is a typical type of linear regression model which describes the linear relationship between a response variable *Y* and an explanatory variable *X*. In a typical regression model, the relationship between variables can be expressed as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \quad (2.2.1)$$

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1j} & \dots & X_{1p} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & X_{i1} & X_{i2} & \dots & X_{ij} & \dots & X_{ip} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nj} & \dots & X_{np} \end{bmatrix} \quad (2.2.2)$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \dots \\ b_p \end{bmatrix} \quad (2.2.3)$$

$$\mathbf{e} = \begin{bmatrix} e_1 \\ \dots \\ e_n \end{bmatrix} \quad (2.2.4)$$

where p is the number of explanatory variables. n is the number of observations. $\boldsymbol{y}$ is the vector of response variable. $\boldsymbol{X}$ is the matrix of explanatory variable. $\boldsymbol{b}$ is the vector of coefficient. $\boldsymbol{e}$ is the vector of residue. Using least square method, we can estimate $\boldsymbol{b}$:

$$\mathbf{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad (2.2.5)$$

2.2.1 ANOVA

In GLMs, the variance of the response variable can be caused by treatment and random effects. We assume that random effects can be expressed as:

$$\mathrm{e}_j \sim \mathrm{N}(0, \sigma_e^2) \quad (2.2.6)$$

where σ_e is standard deviation of random errors. We also assume that the effects of treatments are fixes, then the response variable Y_{ij} is normally distributed:

$$\mathrm{Y}_{ij} \sim \mathrm{N}(\mathrm{T}_i, \sigma_e^2) \quad (2.2.7)$$

where T_i represents treatment effects. Under these assumptions, we can calculate sum of squares that reflects the total variance of the response variable:

$$\mathrm{SSTO} = \sum_i \sum_j Y_{ij}^2 \quad (2.2.8)$$

The total variance that is caused by random errors can be calculate by the following equation:

$$\mathrm{SSE} = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = \sum_i \sum_j \left(Y_{ij} - \widehat{Y_{ij}}\right)^2 \quad (2.2.9)$$

The variance that is caused by treatment effects can be calculated by:

$$\mathrm{SST} = \mathrm{SSTO} - \mathrm{SSE} \quad (2.2.10)$$

To test the significance of GLMs, we use the *F*-statistic:

$$\mathrm{F} = \frac{SSR/p}{SSE/(n-p-1)} \sim F(\mathrm{p}, \mathrm{n}-\mathrm{p}-1) \quad (2.2.11)$$

Using similar methods, we can calculate percentage of variable that can be explained by each explanatory variable. The sum of percentage of variables that can be explained by each variable is defined as R^2 (coefficient of multiple determination). Adjusted coefficient of multiple determination is also calculated in `Expmeasure`. It can be expressed as:

$$\mathrm{R}_{adj}^2 = 1 - \frac{SSE/(n-p-1)}{SSTO/(n-1)} \quad (2.2.12)$$

2.2.2 Multiple comparisons

To test the significance of the difference of the response variable across various groups, comparisons between each pair of treatment should be performed. For t treatment (τ_1, τ_2, …, τ_t) in a single variable, we donate n_i as the number of observations for each treatment τ_i. The total number of observations in the experiments is $N = n_1 + n_2 + \ldots + n_t$.

For each comparison, the null hypothesis is H_0: $\tau_i = \tau_j$ and the alternative hypothesis H_1: $\tau_i \neq \tau_j$. Four types of comparison methods are considered in `Expmeasure`:

(1) LSD test

In LSD test, `Expmeasure` calculates t statistics which can be expressed by:

$$\mathrm{t}^* = \frac{\overline{Y_{i\cdot}} - \overline{Y_{j\cdot}}}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}} \sim t(N-t) \quad (2.2.13)$$

where MSE is the mean squared error value, which equals to the denominator of F statistic described in eq 2.2.11, $\overline{Y_{i\cdot}}$ is the average value of the vector of response variable for treatment τ_i. The corresponding p value is calculated based on the t statistics.

(2) Tukey's test

Test statistic in Tukey's test can be expressed as：

$$\mathrm{q}^* = \frac{\max(\overline{Y_{i\cdot}}) - \min(\overline{Y_{j\cdot}})}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}} \sim q(t, N-t) \quad (2.2.14)$$

All parameters have the same meaning as eq 2.2.13. Similar to LSD test, corresponding p value is calculated based on q statistics.

(3) Scheffé's test

Test statistic in Scheffé's test can be expressed as：

$$\widehat{\mathrm{L}_u} = \sum_{i=1}^{t} c_{iu} \overline{Y_{i\cdot}}, \ and \ \sum_{i=1}^{t} c_{iu} = 0, \ \ u = 1,2,\ldots,m \quad (2.2.15)$$

The contrast is arbitrarily defined and can contain an infinite number of comparisons. In this case, simultaneously confidence intervals in that the probability is at lest 1-α that all of them are correct can be expressed as:

$$\widehat{\mathrm{L}_u} - S_u \cdot \sqrt{MSE \cdot \sum_{i=1}^{t} \frac{{c_{iu}}^2}{n_i}} < \mathrm{L}_u < \widehat{\mathrm{L}_u} + S_u \cdot \sqrt{MSE \cdot \sum_{i=1}^{t} \frac{{c_{iu}}^2}{n_i}} \quad (2.2.16)$$

$$where\ S_u = \sqrt{(t-1)F(\alpha; t-1, n_T - t)} \quad (2.2.17)$$

(4) Bonferroni test

It is similar to Scheffé's test. The test statistic for *k* comparisons can be expressed as:

$$\mathrm{L} = \sum_i c_i \tau_i \quad (2.2.18)$$

The corresponding 1-α confidence interval is:

$$\hat{\mathrm{L}} \pm \mathrm{t}\left(\frac{\alpha}{2k}, n_T - t\right) \sqrt{MSE \cdot \sum_{i=1}^{t} \frac{{c_u}^2}{n_i}} \quad (2.2.19)$$

2.2.3 Residual plot

In regression analysis, an important postulation is normal distribution of random noise (See eq 2.2.6). Consequently, residues should not exhibit significant deviation from the normal curve.

To test this, we usually make a kernel density estimation of residue from its histogram. If the estimated curve shows a different pattern from the normal curve, we believe that the regression model is not reliable. An alternative method is Normal Q-Q plot (Normal Quantile-Quantile plot), which is defined as the probability of studentized residue verses corresponding quantile of standard normal distribution. We expect that each point on the plot should not significantly deviate from *y=x* line. Otherwise, the regression model may be inaccurate.

2.2.4 Jackknife test

In cross-validation techniques, samples that we are interested in are divided into *k* subsets. In each step, one of the subsets is chosen as training dataset to construct regression models. R^2 coefficient is calculated for each regression model before assessing its change over *k* repeats. This method is also known as jackknifing.

2.2.5 Detect outliers

Outliers, high leverage points, and high influence points may cause great negative effects to regression models. A sensible way is to delete them before make conclusions based on regression results.

Outliers are those points cannot be predicted by regression model. This is often indicated by high absolute value of residue. An arbitrary rule to identify outliers is to points with studentized residues higher than 2 or lower than –2.

High leverage points are highly linked to other explanatory variables, which can be identified using hat statistic. Observations with a hat value higher than 2 (or 3) fold of average hat value (p/n, where p is the number of coefficients in the regression model and n is the number of observations) can be regarded as high leverage points.

High influence points have a great impact on the coefficients in the model, which can be identified using Cook's value. Observations with a Cook's value higher than $4/(n-k-1)$ (where n is the number of observations and k is the number of explanatory variables) is high influence points.

2.3 *K*-function

Test of overall trend by Pearson coefficient, T-test, and ANOVA is fine. Unfortunately, these techniques do not tell us to what extent the responsible variable is affected by our experimental design.

In geostatistics, a standard tool for measuring dependence of each observation is called *correlation*: regularity, independence, and clustering. The empirical K-function, a typical estimator of correlation, calculates the average number of neighbors for each data point in two-dimensional space. The K-function can be expressed as:

$$\mathrm{K(r)} = \frac{|W|}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{1}\{d_{ij} \leq r\} e_{ij}(r) \qquad (2.3.1)$$

Where $e_{ij}(r)$ is an edge correction weight. $|W|$ is the investigated area in two-dimensional space. n is the number of data points. When the distance between point i and point j (d_{ij}) is smaller than scale parameter r, $\mathbf{1}\{d_{ij} \leq r\} = 1$. Otherwise, $\mathbf{1}\{d_{ij} \leq r\} = 0$. Higher K value reflects a clustering pattern of data points.

For each indicator of transcription or protein production, the value locates in one-dimensional space. Variation of data points is caused by both treatment effects and random error causes. The latter includes the uncertainties of the concentration of reagents, error in incubation time, fluctuation of temperature, etc. Variation caused by linear treatment effects is often larger than that by random effects, which results in an aggregated pattern at small scales but regular pattern at larger scales (Fig. 2.3.2).

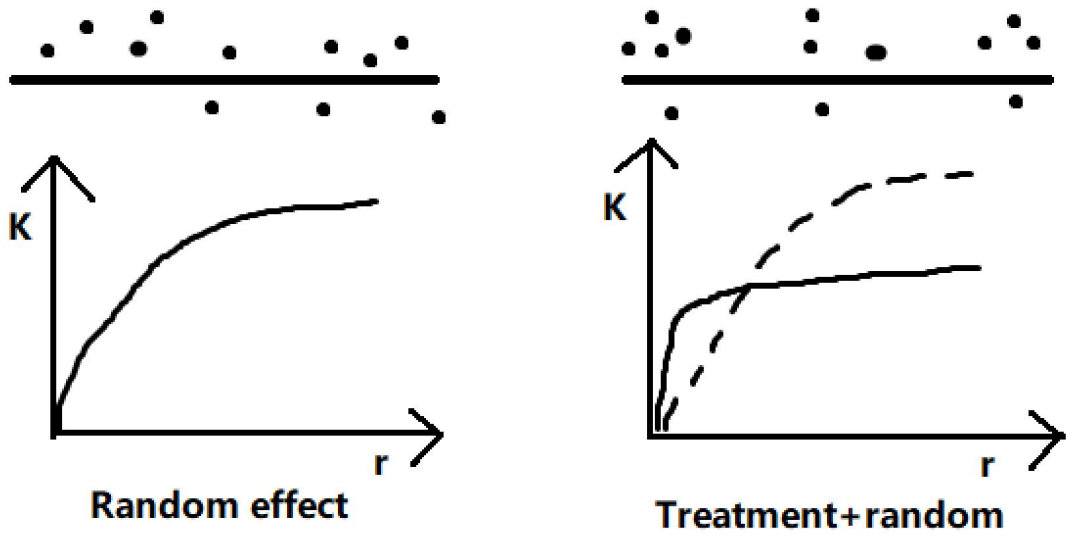

Figure 2.3.2. Point pattern in one-dimensional space caused by random effect or treatment + random effects. A clustering pattern can be detected after adding linear treatment effect.

Consequently, a similar *K*-function can be used as an estimator of the intensity of first-order effect caused by treatment. The *K*-function can be expressed as:

$$\mathrm{K(r)} = \frac{|L|}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathbf{1}\{d_{ij} \leq r\} e_{ij}(r) \quad (2.3.2)$$

Where $|L|$ is the range of the response variable, namely min(X) and max(X) (X is the vector of observed data). Other parameters have been described in equation (2.3.1).

Unfortunately, the *K*-function is incomparable for different experiments because random errors may not be constant. Standardizing K value based on random error is thus needed. Assume that the random error follows a normal distribution $N(\mu_0, \sigma_0^2)$, then theoretical K value for random error is

$$\mathrm{K_0(r)} = |L| E\left(E\left(\mathbf{1}\{d_{ij} \leq r\}\right)\right) = |L| \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \mathbf{1}\{d_{ij} \leq r\} p(j) dj \right) p(i) di \quad (2.3.3)$$

$$= \int_{-\infty}^{+\infty} \left(\int_{i-r}^{i+r} p(j) dj \right) p(i)\, di \quad (2.3.4)$$

where

$$\mathrm{p(i)} = \frac{1}{\sqrt{2\pi}\sigma_0} exp\left(-\frac{(i-\mu_0)^2}{2{\sigma_0}^2} \right) \quad (2.3.5)$$

Then we obtain the relative K value (RK) which reflects the intensity of first-order treatment effect. It can be expressed as:

$$\mathrm{RK(r)} = \frac{K(r)}{K_0(r)} \quad (2.3.6)$$

Although a one-dimensional space is infinite, dataset from measurement result is not. Consequently, an edge correction method is needed for the calibration of K value. When calculating $K(r)$ for a particular scale r, only cases that lies within a range of [min(X) + r, max(X) – r] is considered. This technique is referred as "border correction" [8]. After that, we

obtain edge correction weight $e_{ij}(r)$:

$$e_{ij}(r) = \begin{cases} 1, & if \min(X) + r < i < \max(X) - r \\ 0, & Otherwise \end{cases} \quad (2.3.7)$$

Although higher-order effect may reduce *K* value, we can restrict attention to a smaller range of treatment effect. Imagine that fluorescence activity increases when incubation time rises from 1h to 6h, but further decrement is detected from 6h to 12h. It is a good practice to report a *K* value with a given range 1–6h, or even smaller. We can even measure the sensitivity of a part to a particular explanatory variable using the following equation:

$$S(r) = \frac{\Delta RK(r)}{\Delta v} \quad (2.3.8)$$

where Δv donates the change of explanatory variable.

Low *K* value can be caused by weak higher-order effect, treatment effect, and large sampling interval. Although higher-order effect can be detected by restrict the range of explanatory variable, it is still hard to determine whether a "weak trend" is actually result from a rough experimental design, say (0, 6, 12, 18, 24) h instead of (0, 1, 2, 3, 4, …, 24) h. In order to address this problem, I introduce a second statistic – Treatment-to-error Ratio (*TNE*).

2.4 Spatial interpolation

Spatial interpolation is a method that can predict unknown values based on sample data. In geostatistics, we assume that everything is correlated with each other in two-dimensional space. Objects that are closed to each other share more similarities. In part characterization work, part behavior can also be described by two treatments, which are corresponding to *x*- and *y*-axis in geographical expressions. Two-way experiments are often used in part characterization. However, our design could not cover all levels of treatments. Interpolation is thus helpful for us to identify part behavior in 'treatment region' that is not sampled in the experiment.

Various interpolation techniques are available in geostatistics. I only choose a couple of typical methods in `Expmeasure`. This includes trend, spline, inverse, simple moving average, and kriging interpolation.

2.4.1 Trend surface analysis

Trend surface analysis is actually based on polynomial model. Parameters in the model are estimated using least squares method. When the maximum order is 1, the fitted model can be expressed as:

$$Y(t_i, s_j) = b_0 + b_1 t_i + b_2 s_j \quad (2.4.1)$$

where *t*, *s* is the treatment variables. *Y* is the value of response variable. When the maximum order becomes 2, the model can be expressed as:

$$Y(t_i, s_j) = b_0 + b_1 t_i + b_2 s_j + b_3 t_i^2 + b_4 s_j^2 + b_5 t_i s_j \quad (2.4.2)$$

Once the model is constructed, the response variable can be predicted in each region of "treatment space".

2.4.2 Spline interpolation

Similar to trend surface technique, this method is also based on polynomial model. However, the interpolant is a special type of piecewise polynomial called a spline function. The results of interpolation meet the requirement of the minimum surface curvature of the input data. Although high degree polynomials can be used to generate high-order surface, we usually use 1- or 2-order spline function. We are more likely to get a smooth surface in this method compared to other interpolation techniques.

2.4.3 Inverse distance weight (IDW)

In IDW, predicted points that are closed to observed points can be highly affected by the observed values. This principle is based on the relationship between the similarities of two objects and their distance. Thus, the predicted values can be expressed as:

$$z^*(x_0) = \sum_{i=1}^{n} \lambda_i z(x_i) \quad (2.4.3)$$

where $z^*(x_0)$ is the predicted value at point x_0. $z(x_i)$ is observed value at point x_i. n represents the number of observations that are closed to x_0. λ_i is a weight coefficient that determines the importance of each observed values, which can be expressed as:

$$\lambda_i = \frac{d_{i0}^{-p}}{\sum_{i=1}^{N} d_{i0}^{-p}}, \quad \sum_{i=1}^{n} \lambda_i = 1 \quad (2.4.4)$$

where d_{i0} is the distance between predicted point x_0 and observed point x_i. The power index p measures the impact to the predicted value caused by each observed value. When $p = 0$, no observed points will cause any impacts on prediction.

2.4.4 Simple moving average (SMA)

In SMA, the predicted value in any point x_0 is only determined by the observed points that has a distance short than R from x_0. In other words, only observed points that locates in a "sampling window" can affects the predicted values at x_0. Unlike IDW, the values in x_0 is simply the averaged candidate observed values. Consequently, discrete surface may occur in SMA.

2.4.5 Kriging

As one of the most popular and commonly used interpolation methods, kriging calculates weight coefficient for each observed value, which is similar to IDW. Weight coefficient in kriging is determined by semivariogram. Kriging can be further divided into multiple methods. In `Expmeasure`, only ordinary kriging is available. Parameters in semivariogram is automatically estimated although it is better determined manually in geographical research.

2.5 Trend-to-Error statistic

Signal-to-noise ratio (SNR or S/N) is widely used measure in science and engineering that characterize the intensity of a desired signal compare to background noise. In measurement practices, variance caused by treatment effect should also be compared with that caused by random error.

We define the *TNE* statistic as:

$$TNE = \frac{\sigma_t^2}{\sigma_0^2} \quad (8)$$

Where σ_0^2 donates variation caused by "nugget random error". σ_t^2 is the variation caused by both "nugget" and treatment-related random error.

The inspiration comes from geostatistical methods. According to Tobler's first law of geography, nearby things are more related than distant things. Particularly, observed values in identical place should be the same. But this is not true in the real world due to pure random variation caused by sampling error, which is known as "nugget effect" [11]. A popular statistic – variogram $\gamma(h)$ – demonstrates how variation can be caused by both nugget effect and distance-related effect (Fig. 2.5.1).

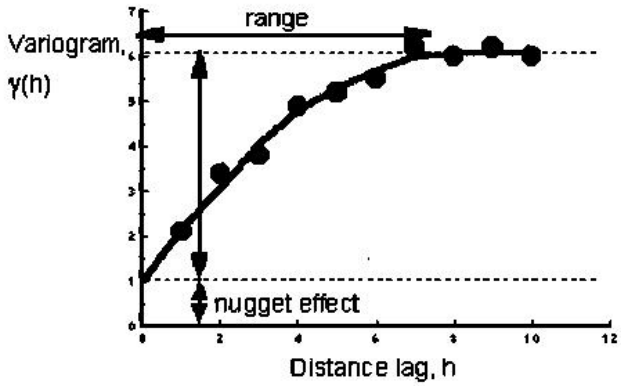

Figure 2.5.1. A typical trend of variogram. The *x*-axis reflects the distance between two interested points, while the value in *y*-axis is negatively correlated with the correlation between two points. Even if $h = 0$, observed value is not constant, which is called "nugget effect". When $h >$ range, two points are no longer correlated with each other.

Assume that we are interested in the relationship between GFP production and the concentration of molecular induction. In control group, variation of the intensity of fluorescence activity can be result from instrumental, environmental, and observational error. Fluctuation of the induction concentration is added into variance source in experimental group, which is expected to generate additional fluctuation in the dataset (Fig. 2.5.2). Strong difference of variation between control group (nugget effect) and experimental group (nugget + treatment) indicates a significant tend in part kinetics.

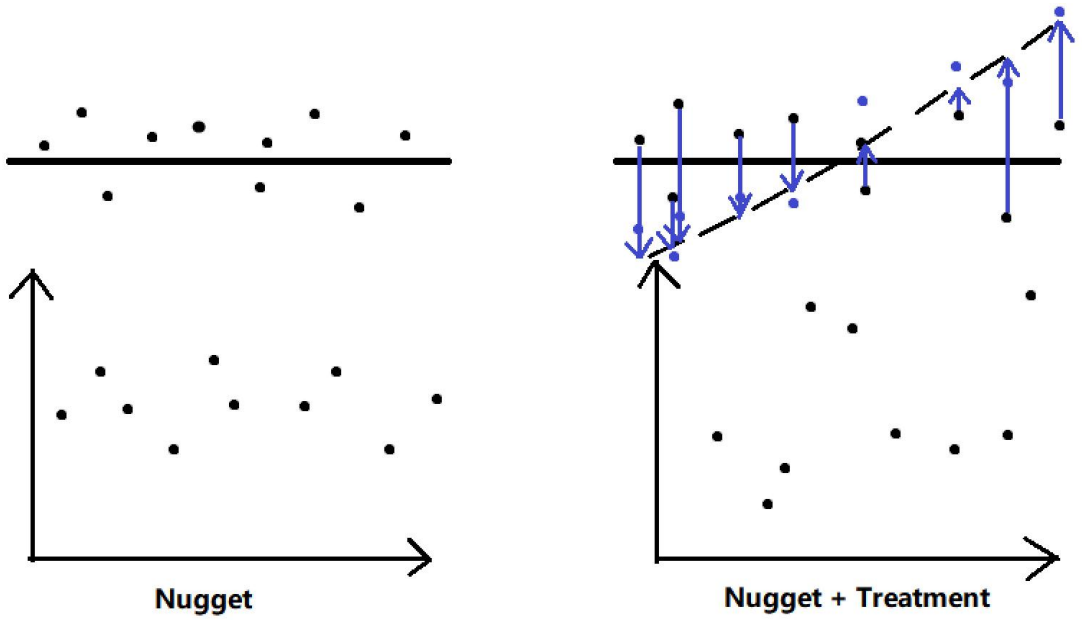

Figure 2.5.2. Variance distribution under the scenario of nugget effect and nugget + treatment effect.

We recommended that a finer-scale experimental design should be considered (e.g. (0, 25, 50, 100) μM rather than (0, 500, 1000) μM) when *TNE* is high.

2.6 T/E Simulation

Simulation is important to assess the uncertainty and reproductivity of statistical inferences. For each wet experiment, variation of the response variable is assumed to be affected by treatment and random effects. Similar to the design of *K*-function and *TNE* index, I assume random effect follows normal distribution $N(0, \sigma_0^2)$. Thus, the response variable Y_i can be expressed as:

$$Y_i = T_i + N\left(0, \sigma_0^2\right) \sim N(T_i, \sigma_0^2) \quad (2.6.1)$$

where T_i is the fixed effect caused by treatments. σ_0^2 represents random effects caused by measurement error. In real experiment, treatment (e.g. concentration of reagents, incubation time, temperature, etc.) may not be stable either. However, as sample size increases, these errors converge to zero. Base on this, we can estimate Y_i:

$$\overline{Y_{i[i\in(k_1,k_2,\ldots,k_n)]}} = \frac{\sum_{T_1=t_{1k_1},T_2=t_{2k_2},\ldots,T_n=t_{nk_n}} T_i}{l_{i\in(k_1,k_2,\ldots,k_n)}} + N\left(0, S_0^2\right) \quad (2.6.2)$$

$$S_0 = \sqrt{\frac{\sum_{T_1=T_2=\ldots=T_n=0}\left(T_i - \frac{\sum_{T_1=T_2=\ldots=T_n=0} T_i}{l_{i\in(0,0,\ldots,0)}}\right)^2}{l_{i\in(k_1,k_2,\ldots,k_n)}}} \quad (2.6.3)$$

where T_1 (ranges from level t_{11}, …, t_{1k_1}, …, t_{1s_1}), T_2 (ranges from level t_{21}, …, t_{2k_2}, …, t_{2s_1}), …, T_j, …, T_n represents treatment effects of variable 1, 2, …, n. *l* represents number of observations for each treatment.

With well-defined linear model, Monte-Carlo simulations can be performed to generate hypothetical experimental data. Regression and ANOVA test can be performed based on simulated data to assess the uncertainty of the percentage of variance that can be explained by each variable.

3 Background

The aim of `Expmeasure` is to

- Provide comprehensive statistical tools for part characterization in iGEM, especially for those who find it challenging to understand principles of statistical or code in the computer.
- Provide novel statistics developed by ZJU-China 2021 that is used for improvement of wet experiments.

Many iGEMers get trouble with choices of a appropriate statistics, mathematical calculation, and data visualization. Software like Origin, SPSS, and SAS can help them to make several basic plots. Unfortunately, advanced methods that is important for redesign the experiment is not available in those tools. Coding is an alternative strategy, but also labor-demanding. `Expmeasure` simply provides a GUI interface based on R shiny, which enables all iGEMers to perform advanced statistical methods (e.g. detect outliers, assess errors, evaluate effects of each explanatory variables, make predictions, and examine variance patterns)

All iGEMers can make robust conclusions and get suggestions of experimental design by simply clicking a couple of icons in `Expmeasure`.

4 Data Input

The DATA page is the only one that requires you to upload your data. To avoid unexpected errors, you should follow the standard format used in `Expmeasure`.

- Only ".csv" file is allowed.
- Data should be database-like style. It means that each row represents an observation, and each column represents a variable (Table 4.1). Some users prefer data table where each row reflects a treatment and each column reflects another treatment (Table 4.2). Please NOTE that such format is NOT allowed in `Expmeasure`.
- **Do NOT present categorical variable using numbers**. This would cause incorrect results rather than any error or warning messages. So do use numbers for continuous variables and characters for categorical variables.

Table 4.1. Format that is accepted in `Expmeasure`.

Var1	Var2	Var3	Var4	Var5	…
value11	value21	value31	value41	value51	…
value12	value22	value32	value42	value52	…
…	…	…	….	…	…

Table 4.2. Format that is NOT allowed in `Expmeasure`.

	Var1	Var2	Var3	Var4	…
Var5	value15	value25	value35	value45	…
Var6	value16	value26	value36	value46	…
…	…	…	….	…	…

Example data is shown in Section 1.1. You can also make your data looks like the test data in order to avoid errors.

You will be required to assign a response variable and several explanatory variables in DATA page. Expmeasure recognizes variables by the column name of your data file (e.g. Var1, Var2, …, Var5 in Table 4.1).

5 Running `expmeasure()`

5.1 Trend

This module provides visualization tools to explore trend (or patterns) in the dataset.

5.1.1 2D Plot

Generate a simple scatter plot with fitted curve using polynomial model. 95% confidence interval is also shown in the plot.

Input
"Select explanatory variable": Specify a candidate explanatory variable which is corresponding to the *x*-axis of 2D Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

Output
A 2D Plot that shows the relationship of the response variable and a single explanatory variable.
x-axis: a candidate explanatory variable that is specified by "Select explanatory variable" in TREND | 2D Plot page.
y-axis: response variable that is specified by picker box "Select response variable" in DATA page.

5.1.2 3D Plot

Generate a 3D scatter plot the fitted response surface using polynomial model. 3D Plot is generated in a separated window (named by "RGL Device") rather than in the TREND | 3D Plot page. You can drag to rotate the picture. You can also scroll the mouse wheel to resize the picture.

Input
"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of 3D Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of 3D Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted polynomial model. Stepwise regression will be carried out before make a 3D plot, so selecting higher-order model may still generate linear response surface because higher-order terms may be deleted during stepwise calculation.

Output

A 3D Plot that shows the relationship of the response variable and two candidate explanatory variables. Each dot in the picture represents an observation. The response surface is generated by regression analyse.

x-axis, *y*-axis: candidate explanatory variables specified by "Select explanatory variable" in TREND | 3D Plot page.

z-axis: response variable that is specified by picker box "Select response variable" in DATA page.

Error messages:

#1 Error: Var# is deleted in stepwise regression. Can not make 3D Plot.

This indicates that after stepwise regression, variable 1 (or variable 2) is deleted because it does not cause significant effects on the response variable. One of the solutions is increasing the order of the polynomial model, which can be achieved by clicking radio buttons (named by "Order") with higher values. If the error message persists, I do not recommend you to make a 3D Plot. Make a 2D Plot without variable 1 (or variable 2) instead.

#2 Error: Non-numeric argument to binary operator

(in Chinese: 二进列运算符中有非数值参数)

This indicates that one of the variables you selected is a categorical variable rather than a numerical one. Or you include some characters in the variable accidentally even if you think it should be a numerical variable. If you want to use categorical variable in 3D Plot, DO NOT choose higher order (2 or 3) because no explanations can be made for a quadratic form of characters.

5.1.3 VioPlot

Generate a violin plot that reflects the relationship between a response variable and a categorical variable.

Input

"Select explanatory variable": Specify a **categorical** explanatory variable which is corresponding to the *x*-axis of the violin plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

Output

A violin plot that reflects the relationship between a response variable and a categorical variable. In the middle of the "violins", there are a couple of box plots that looks like a rectangle with lines extending from the top and bottom.

In the box plot, the top line of the rectangle indicates the third quartile (Q3), while the bottom of the rectangle indicates the first quartile (Q1). Median value is represented by the horizontal line near the middle of the rectangle. A vertical line extends from the rectangle indicates the maximum and minimum values of the variable. Red dots in the plot indicates

outliers while normal observations are represented by dots smaller than outliers. The density curve outside the box represents the frequency distribution of the response variable for each group.

x-axis: the categorical explanatory variable that is specified by "Select explanatory variable" in TREND | VioPlot page.
y-axis: response variable that is specified by picker box "Select response variable" in DATA page.

Error messages:
#1 Error: Continuous x aesthetic -- did you forget aes(group=...)?
This error indicates that you do not input a categorical variable. Ensure that each level in this variable is represented by characters rather than numbers in the raw data.

5.2 Error

This module helps you to find outliers, high leverage points, and high influence points. You may delete outliers before make conclusions about the patterns in your dataset.

5.2.1 Outlier

Generate a table that gives information about outliers in your dataset.

Input
"Select variable 1": Specify a candidate explanatory variable which is used to construct polynomial model. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify a candidate explanatory variable which is used to construct polynomial model. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted polynomial model. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.

Output
A table that provides information about outliers based on polynomial model constructed by stepwise regression.

Columns:
id: No. of observation. This number is corresponding to the No. of row in the raw data (exclude the title row in raw data).
rstudent: Standardized residues from the polynomial model.
p: p-value of outlier test. p value smaller than 0.05 indicates significance of outliers.

bonf.p: p-value based on Bonferroni outlier test.

Error messages:

#1 Error: Non-numeric argument to binary operator

(in Chinese: 二进列运算符中有非数值参数)

This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some characters in the raw data.

5.2.2 Leverage

Generate a table that gives information about high leverage points in your dataset.

Input

"Select variable 1": Specify a candidate explanatory variable which is used to construct polynomial model. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2 (Or Not)": Specify a candidate explanatory variable which is used to construct polynomial model. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Order": The maximum order of fitted polynomial model. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.

Output

A table that provides information about high leverage points based on polynomial model constructed by stepwise regression.

Columns:

id: No. of observation. This number is corresponding to the No. of row in the raw data (exclude the title row in raw data).

hat: The hat value for each high leverage points. For detailed statistical explanations, see Chapter 2.

Error messages:

#1 Error: Non-numeric argument to binary operator

(in Chinese: 二进列运算符中有非数值参数)

This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some

characters in the raw data.

5.2.3 Influence

Generate a table that gives information about high influence points in your dataset.

Input

"Select variable 1": Specify a candidate explanatory variable which is used to construct polynomial model. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2 (Or Not)": Specify a candidate explanatory variable which is used to construct polynomial model. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Order": The maximum order of fitted polynomial model. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.

Output

A table that provides information about high influence points based on polynomial model constructed by stepwise regression.

Columns:

id: No. of observation. This number is corresponding to the No. of row in the raw data (exclude the title row in raw data).

cooks: The Cooks value for each high influence points. For detailed statistical explanations, see Chapter 2.

Error messages:

#1 Error: Non-numeric argument to binary operator

(in Chinese: 二进列运算符中有非数值参数)

This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some characters in the raw data.

5.3 Significance

This module provides a wide range of statistical tools to test the significance of the patterns in your dataset.

5.3.1 *K*-function

Generate *K*-function plot based on experimental simulated data. Monte-Carlo simulations are used to generate confidence envelope.
NOTE: Calculation might be slow. See R/RStudio console to monitor the progress of calculation.

Input
"Parameter *nsim*": The number of Monte-Carlo simulations that generates the grey envelope in *K*-function plot. Cautions should be made that a high value of *nsim* can slow down the calculation. In R/Rstudio console, you can monitor the progress of Monte-Carlo simulations by inspecting the number of simulations that is finished before the *K*-function plot is generated.

"Select variable 1": Specify a numerical explanatory variable for one-way *K*-function plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select a range": Specify a range of the selected numerical variable for one-way *K*-function plot.

Output
An overall *K*-function plot tests the significance of first-order effects in the dataset. An one-way *K*-function plot reflects local significance of the first-order treatment effects of a specific variable. The black line reflects the observable *K* values based on real data for each scale *r*, while the grey envelope reflects the $nsim/(nsim+1) \times 100\%$ confidence interval generated by Monte-Carlo simulations.

x-axis: scale *r* in one-dimensional space. The unit is consistent with the response variable.
y-axis: *K* value.

How to interpret: If the black line is not in the envelope, there is significant first-order effect in the dataset, otherwise no evidence suggests the effect. In particular, the black line should be much higher than the upper bound of the envelope at small scales and be much lower than the lower bound of the envelope at larger scales if the first-order effect is strong. For more details about statistical theory, see chapter 2.

5.3.2 GLMs

Construct generalized linear models (GLMs) and perform *t*-test to examine the significance of effects caused by each covariate to the response variable.

Input

"Select variable 1": Specify a candidate explanatory variable which is used to construct GLMs. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify a candidate explanatory variable which is used to construct GLMs. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted GLMs. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.

Output
A table that gives results of stepwise regression and *t*-test for each variable. Each row represents a term in the model, such as intercept, variable 1, and (variable 1)^2.

For each column:
Estimate: The slope of an item in GLMs.
Std. Error: The standard error of an item in GLMs.
t value: Result of *t*-test for an item.
Pr(>|t|): *p* value of *t*-test.

The output also reports R^2 and adjusted R^2 to measure how well the regression predicts the real data points.

Error messages:
#1 Error: Non-numeric argument to binary operator
(in Chinese: 二进列运算符中有非数值参数)
This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some characters in the raw data.

5.3.3 Residue

Report residues after stepwise regression and examine whether they follow the normal distribution.

Input
"Select variable 1": Specify a candidate explanatory variable which is used to construct GLMs. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify a candidate explanatory variable which is used to construct GLMs. However, you may only have one explanatory variable in your dataset, so

variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted GLMs. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.
"Default": Types of residual plot. If you select "Distribution", a histogram of residues will be generated. If "Q-Q Plot" is selected, the residue will be compared to the quantile of normal distribution in different axis. If "Table" is selected, the output area will simply generate a residual table for each observation.

Output

1. Distribution Plot

A histogram of studentized residual plot will be generated. The yellow curve represents theoretical density of normal distribution. The green curve represents observed residual distribution estimated using kernel methods.

2. Q-Q Plot

x-axis: Quantiles of theoretical normal distribution.
y-axis: Quantiles of observed residues.
The envelope represents confidence interval of quantile-based normal distribution test.

3. Table

id: No. of observation. This number is corresponding to the No. of row in the raw data (exclude the title row in raw data).
residue: The residue for each observation under fitted model.
rstudent: Studentized residue for each observation under fitted model.

Error messages:

#1 Error: Non-numeric argument to binary operator
(in Chinese: 二进列运算符中有非数值参数)
This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some characters in the raw data.

5.3.4 Jackknife

Perform Jackknife tests to assess the robustness of observed relationship.

Input

"Select variable 1": Specify a numerical explanatory variable which is used to construct GLMs. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2 (Or Not)": Specify a numerical explanatory variable which is used to construct GLMs. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted GLMs. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.
"Parameter *k*": No. of Jackknife tests. With higher *k* you can test the robustness of regression more confidently.

Output
Report value includes original R^2, *k* fold cross-validated R^2, and the change of R^2.

Error messages:
#1 Error: Non-numeric argument to binary operator
(in Chinese: 二进列运算符中有非数值参数)
This indicates that one of the variables you selected is a categorical variable. Do NOT includes categorical variable in this page. Check if there are characters in the raw data.

#2 Error: NA/NaN/Inf in 'x'
The same problem as #1.

5.4 Explanatory

5.4.1 ANOVA

This module tells you about the relative importance of each explanatory variables in the experimental design.

Input
"Select variable 1": Specify an explanatory variable which is used to construct GLMs. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify an explanatory variable which is used to construct GLMs. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted GLMs. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.

Output
An ANOVA table.

items: Each item in fitted GLMs.
df: Degree of freedom for each item.
Sum.Sq: Sum of square for each item.
Mean.Sq: Mean of square for each item.
F.value: Value of *F* test for each item.
P.value: *P* value of F test for each item.
Explained: Percentage of variance that can be explained by each item.

Error messages:
#1 Error: Non-numeric argument to binary operator
(in Chinese: 二进列运算符中有非数值参数)
This indicates that one of the variables you selected is a categorical variable, and you also choose a higher order (2 or 3) to construct the polynomial model. But this does not make sense. Try to replace the variable to a numerical one or reduce order. If you think the input variables are numerical one but the error still occurs, check whether there are some characters in the raw data.

5.4.2 Muti Comp

Perform multiple comparisons for a single variable in GLMs.

Input
"Select variable 1 (Also for multiple comparisons)": Specify an explanatory variable which is used to construct GLMs. This variable is used for multiple comparisons. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify an explanatory variable which is used to construct GLMs. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Order": The maximum order of fitted GLMs. Never choose a higher order (2 or 3) if variable 1 (or variable 2) is a categorical variable.
"Default": Methods of multiple comparisons.

Output
A table that shows the result of multiple comparisons.
"level": All levels of variable 1.
"y.d": The mean value of the response variable for each level.
"groups": Groups with the same mark show no significance treatment effects. For example, in Table 5.4.1, there is no significant difference of fluorescence activity between group "Cu" and "Mn", while the difference of that between group "Mn" and "Co" is dramatic.

Table 5.4.1. An example of the result of multiple comparisons.

level	y.d	groups

Cu	31253.60	a
Mn	30272.06	a
Co	27613.17	b
Zn	26105.06	b
Ni	26058.61	b
CK	15768.27	c

Error messages:

#1 Error: After regression, 'var1' is excluded. Can not make mutiple comparision for 'var1'.

This indicates that variable 1 causes no effect to the response variable and thus being excluded when constructing GLMs using stepwise regression. You can try higher-order effects by increasing order or to avoid adding this variable into the model.

5.5 Prediction

This module provides a wide range of geostatistical techniques to predict two-dimensional kinetics behavior of any biological part.

5.5.1 Trend

Give a prediction of part kinetics using trend analysis.

Input

"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of Trend Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of Trend Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Order": The maximum order of fitted polynomial model. Stepwise regression will be carried out before make a Trend Prediction Plot, so selecting higher-order model may still generate linear response surface because higher-order terms may be deleted during stepwise calculation.

Output

Trend Prediction Plot – a plot that gives the predicted relationship between the response variable and two explanatory variables.

x-axis: Values of variable 1.

y-axis: Values of variable 2.

z-axis: Predicted values of the response variable.

Error messages:

#1 Error: 'var#' is not a continuous variable. Can not make trend prediction.
Replace variable # from a categorical variable to a numerical one. Check if there are any characters in the raw data file.

5.5.2 Spline

Give a prediction of part kinetics using spline interpolation method. For statistical details about spline function, see Chapter 2.

Input
"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of Spline Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of Spline Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

Output
Spline Prediction Plot – a plot that gives the predicted relationship between the response variable and two explanatory variables.

x-axis: Values of variable 1.
y-axis: Values of variable 2.
z-axis: Predicted values of the response variable.

Error messages:
#1 Error: 'var#' is not a continuous variable. Can not make trend prediction.
Replace variable # from a categorical variable to a numerical one. Check if there are any characters in the raw data file.

5.5.3 Inverse

Give a prediction of part kinetics using inverse interpolation method. For statistical details about inverse function, see Chapter 2.

Input
"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of Inverse Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of Inverse Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Inverse distance power *p*": The power index in inverse interpolation methods, which determines the importance of the values of observed points to that of unsampled points.

Output

Inverse Prediction Plot – a plot that gives the predicted relationship between the response variable and two explanatory variables.

x-axis: Values of variable 1.
y-axis: Values of variable 2.
z-axis: Predicted values of the response variable.

Error messages:

#1 Error: 'var#' is not a continuous variable. Can not make trend prediction.

Replace variable # from a categorical variable to a numerical one. Check if there are any characters in the raw data file.

5.5.4 Simp Mov

Give a prediction of part kinetics using simple moving average (SMA) interpolation method. For statistical details about SMA, see Chapter 2.

Input

"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of SMA Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of SMA Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Radius of window *R*": The radius of the interpolation window.

Output

SMA Prediction Plot – a plot that gives the predicted relationship between the response variable and two explanatory variables.

x-axis: Values of variable 1.
y-axis: Values of variable 2.
z-axis: Predicted values of the response variable.

Error messages:

#1 Error: 'var#' is not a continuous variable. Can not make trend prediction.

Replace variable # from a categorical variable to a numerical one. Check if there are any characters in the raw data file.

5.5.5 Kriging

Give a prediction of part kinetics using kriging interpolation method. For statistical details

about kriging, see Chapter 2.

Input

"Select variable 1": Specify a candidate explanatory variable which is corresponding to the *x*-axis of Kriging Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2": Specify a candidate explanatory variable which is corresponding to the *y*-axis of Kriging Prediction Plot. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

Output

Kriging Prediction Plot – a plot that gives the predicted relationship between the response variable and two explanatory variables.

x-axis: Values of variable 1.

y-axis: Values of variable 2.

z-axis: Predicted values of the response variable.

Error messages:

#1 Error: 'var#' is not a continuous variable. Can not make trend prediction.

Replace variable # from a categorical variable to a numerical one. Check if there are any characters in the raw data file.

5.6 Uncertainty

5.6.1 TNE

Calculate *TNE* index for each treatment. For statistical details, see chapter 2.

Input

"Select variable 1": Specify an explanatory variable which is used to divide raw data into different groups based on treatment level for *TNE* calculation. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

"Select variable 2 (Or Not)": Specify an explanatory variable which is used to divide raw data into different groups based on treatment level for *TNE* calculation. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.

Output

A *TNE* plot that visualize the *TNE* value for each group.

x-axis: Values of variable 1.

y-axis: Values of variable 2.
z-axis: Values of *TNE*. (if variable 2 is not in use, it reduces to *y*-axis)

Error messages:
#1 Error: Not enough observations
For each treatment, your dataset only has a single observation. This is dangerous because you cannot make any statistical inference based on data without any repetition in the same treatment.

5.6.2 T/E Simu

Generate T/E Simu plot that reflects the uncertainty of variances that can be explained by candidate variable(s). For statistical details, see chapter 2.

Input
"Select variable 1": Specify an explanatory variable which is used to perform T/E simulations. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Select variable 2 (Or Not)": Specify an explanatory variable which is used to perform T/E simulations. However, you may only have one explanatory variable in your dataset, so variable 2 is optional. Choose "Not Use" options if variable 2 is not available in your dataset. Only variables that are selected in "Select explanatory variables" picker input in the DATA page can be chosen.
"Number of Simulation": Number of simulations.

Output
A T/E Simu plot that demonstrates the dynamics of the percentage of variance that can be explain by one or two variable(s).

x-axis: The number of simulations.
y-axis: Percentage of variance that can be explained by a specific variable.

Error messages:
#1 Error: 0 (non-NA) cases
The same problem as #1 Error in "TNE" panel. For each treatment, your dataset only has a single observation. This is dangerous because you cannot make any statistical inference based on data without any repetition in the same treatment.

6 Suggested Workflow

6.1 Basic Concepts

I suggest a general framework of statistical inference about part kinetics (Fig. 6.1.1):

Step1: Explore possible trends in the dataset.

- If there is only a single numerical variable in your dataset, use 2D Plot.
- If there are two numerical variables in your dataset, try to use 3D Plot to make some predictions about relative importance of each explanatory variable.
- If there are various numerical variables (although I know this occasion is rare in wet lab), try all possible combinations of two variables to make 3D Plot and discard variables that has no influence on the response variable.
- If categorical variable(s) exist in your dataset, make a violin plot to examine whether different levels in this variable can cause difference of the response variable.

Obviously if you identify a trend, you can skip Step 2 and start to test the significance of the trend. But deleting outliers in your dataset before downstream analyses is still worth. If you do not identify a trend, go to Step 2 and delete outliers in your dataset.

Step2: Identify outliers in the dataset.

Outliers, high leverage points, and high influence points are observations that situates away from the main body. The occurrence of these points may due to accidents, and may reduce patterns in the dataset. I suggest you to delete these points in the raw dataset and reload the data to `Expmeasure`.

Step3: Test the significance of observed trend.

- The first thing is calculating overall *K* values. Of *K* function is located within the envelope, try to specify a numerical variable with its range and replot the *K* function. If *K* function still locates within the envelope, there is no significant trend in the dataset and construction of GLMs will be not sensible. So, if this happens, please return to your wet lab work.
- If significant trend is detected in *K* function plot, you can construct GLMs and perform *t*-test for each candidate variable. Results of stepwise regression using GLMs tell you how candidate explanatory variables have an impact on the response variable.
- Make a local *K* function plot by selecting a variable and its range to test the significance of first-order effect. Range selection should be based on the results of GLMs.
- Make a residual plot to test the robustness of GLMs. If residuals do not follow the normal distribution, you can NOT report your GLMs results to future iGEM teams because they may be inaccurate.
- Make a *F*-test to examine the significance of GLMs.
- Make a jackknife test as a complementary method to examine the significance of GLMs.

Confidence of observed trend should be based on a variety of statistical methods. Once you make sure that the trend is significant, you want to explore the relative importance of each

variable.

Step4: Explore the relative importance of each explanatory variable.

- Make an ANOVA test to calculate the percentage of total variance that can be explained by each explanatory variable.
- Make multiple compassions for each candidate variable to test the significance of the response variable between difference levels of the variable.

Step5: Make predictions of part behavior

Trend analysis, Spline interpolation, Inverse interpolation, Simple Move Average, and Kriging are different methods that can extrapolate the values of response variable based on different theories.

Step6: Uncertainty analysis

- Calculate *TNE* to identify observations that may exhibit a significant first-order effect. Further experimental design should consider levels that are closed to points with high *TNE* value.
- Perform T/E simulations to examine the uncertainty of the percentage of variance that can be explained by each variable.

Step7: Communication and Contribution

Return to *K* function plot and report it to other iGEM teams to tell them in which range they can observe significant treatment effects. Report interpolation results to give information to other iGEM teams about part kinetics. Report T/E simulation results to make sure that your results are reliable.

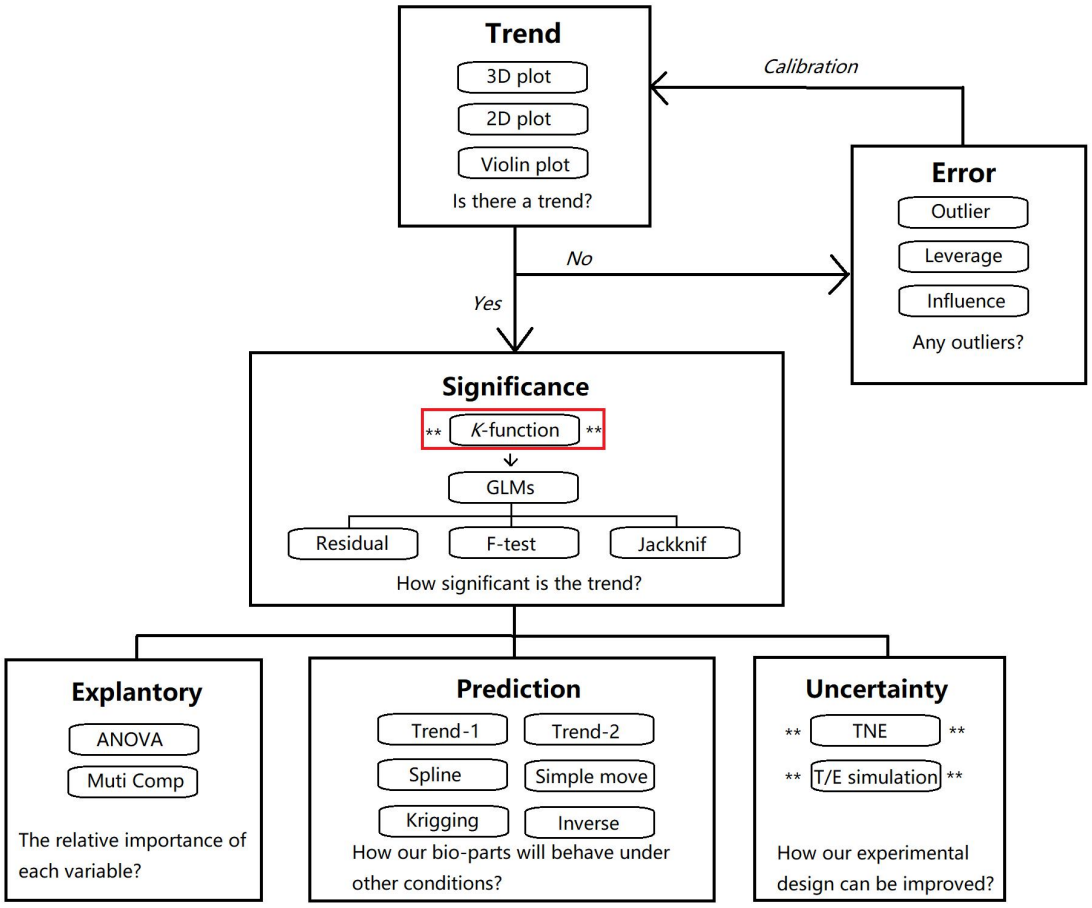

Figure 6.1.1. General protocol of statistical test for part kinetics characterization. Noval statistics in `Expmeasure` are marked by "*".

6.2 Case Study 1: Characterization of part *sPrcn*

Action 1: Pre-experiment

In iGEM competition, part characterization is important for reuse. A well-documented part can be used easily by other iGEM teams without unexpected results. Unfortunately, incomplete analyses may produce incorrect results, which is misleading. This case demonstrates how `Expmeasure` can help us make a sensible conclusion based on our dataset and guide our experiments using the framework described in Section 6.1.1.

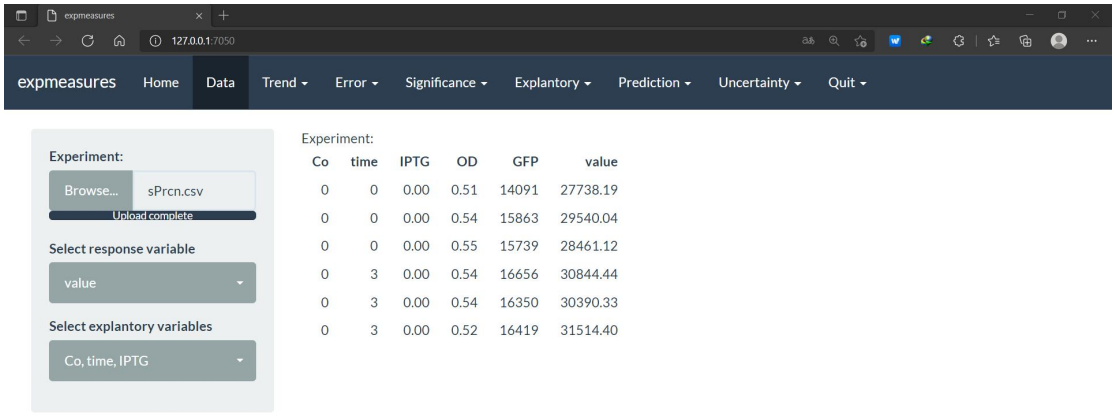

Figure 6.2.1. Upload data `sPrcn.csv`

Upload `sPrcn.csv` on "Data" page. Select "value" as the response variable. Select Co, time, and IPTG as explanatory variables.

We first generate a 2D Plot to identify the relationship between the concentration of Co^{2+}, IPTG, and the fluorescence activity. In the 2D Plot, we find that no significant effects by Co^{2+} can be detected while IPTG can considerably enhance the fluorescence activity (Figure S1, S2). This trend can be further demonstrated by 3D Plot, as shown in Figure 6.2.2. Consequently, **it seems that Co^{2+} can not make any contribute to the fluorescence activity**, which is contradict to our expectation. This is because sP*rcn* is improved from P*rcn*, a cobalt-sensitive promotor. This contraction may be due to outliers, imperfect experimental design, or the fact that P*rcn* is actually not a cobalt-sensitive promotor. If analyses stop at this point, incorrect conclusion will be made.

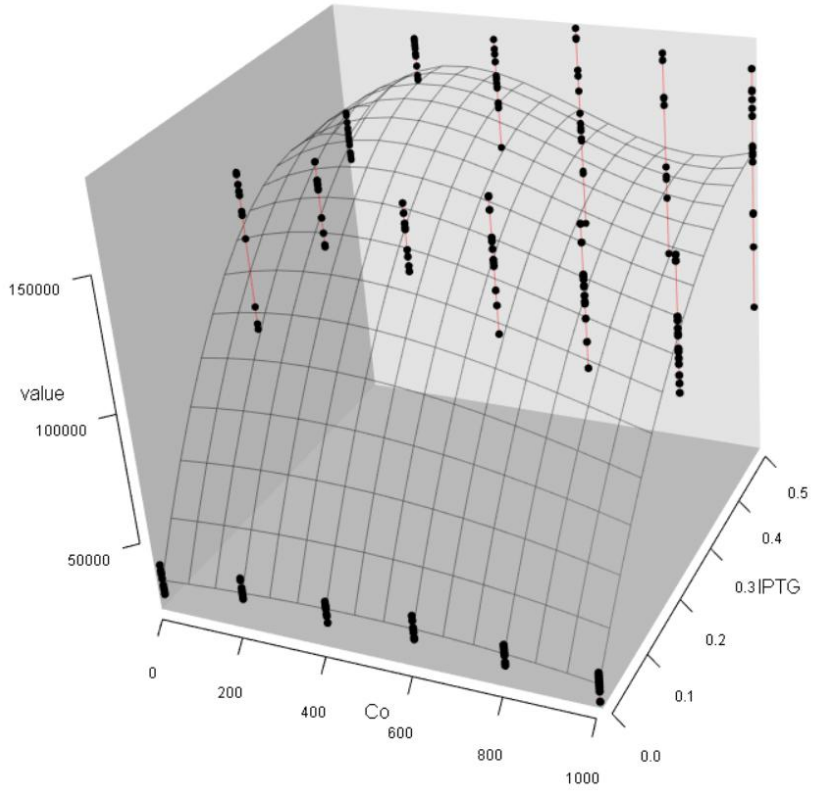

Figure 6.2.2. Relationship between the concentration of Co^{2+} (or IPTG) and the intensity of fluorescence activity.

Turn to the "Error" Page and try to detect outliers in the dataset. However, only one outlier is detected. No leverage point lies in the dataset, while large quantity of high influence points is detected. This indicates that the trend in 3D Plot may be vulnerable to a couple of points in the dataset.

Turn to the "Significance" Page in order to assess whether the trend is strong enough. The overall *K*-function plot indicates a significant trend in the dataset. However, treatment effect of IPTG is significantly larger than that of Co^{2+}, indicated by higher degree of clustering at small scales and higher degree of regular patterns at larger scales (Fig. S3, S4). In GLM analysis, the variable of Co^{2+} concentration is even deleted by stepwise regression when order is set as 1. When GLM model only contains the explanatory variable of Co^{2+}, the residues is severely deviated from the normal distribution (Fig. S5). Only by adding the effect of IPTG, the model makes sense.

In the "Explanatory" Page, we get the information that the effects of IPTG can explain 82.21% variation of the response variable (i.e. the intensity of fluorescence activity per cell). In "Prediction" Page, we still get a similar trend as 3D Plot.

The most important analysis lies in the last page – "Uncertainty" Page, which may provide implications about how to improve our experimental design. It may be useful to address the contradiction mentioned above. In TNE plot, an unexpected high TNE value is detected at

lower Co^{2+} concentration (i.e. (Co^{2+}, IPTG, TNE) = (0μM, 0.25Mm, 8.62), Fig. 6.2.3). Although extremely high concentration of Co^{2+} can also cause high TNE value, we think another fine-scale trend exists in lower Co^{2+} concentration. This is because we found that 1000μM can already kill the most of bacterial cells when we are measuring the fluorescence activity. Consequently, **fine-scale design of lower concentration of Co^{2+} should be taken into consideration**, indicated by high TNE value.

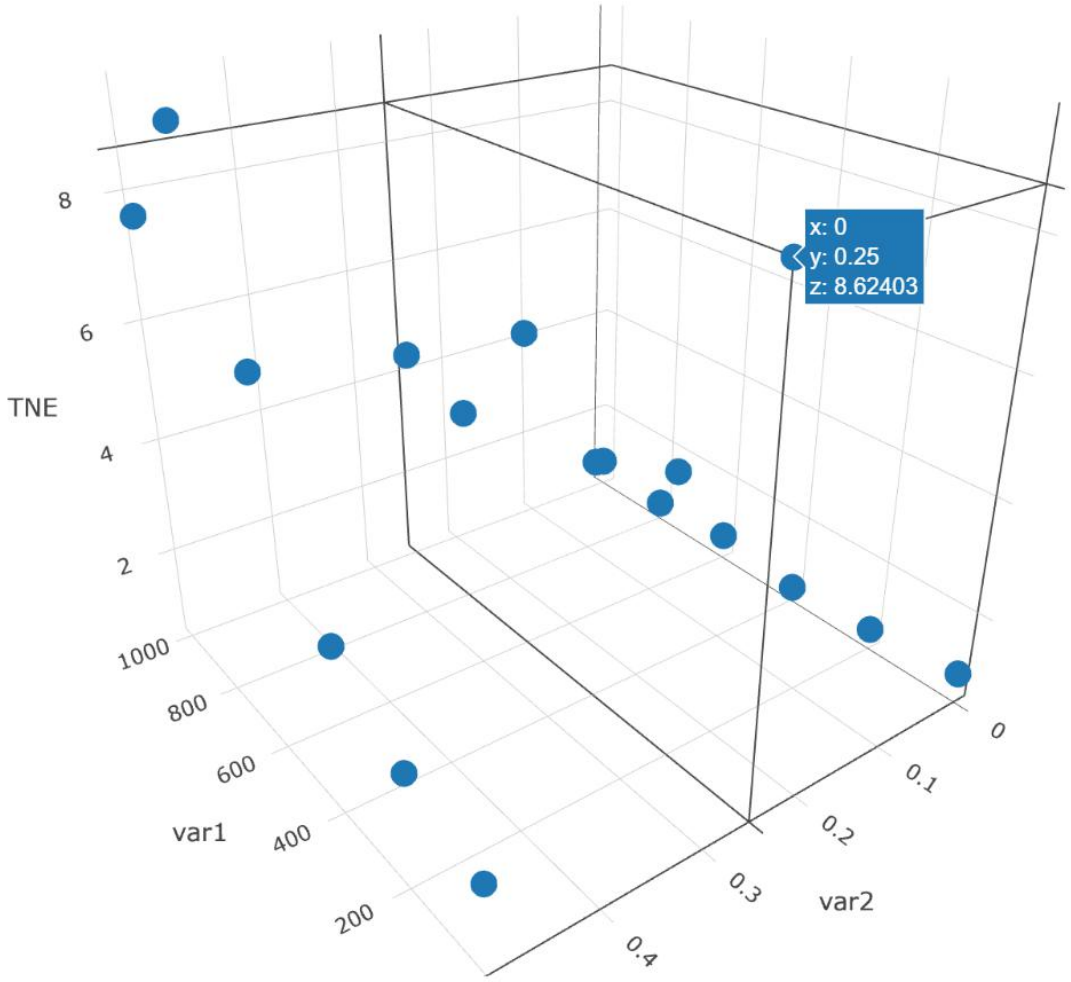

Figure 6.2.3. TNE Plot of Co^{2+} (var1) and IPTG concentration (var2). The *z*-axis reflects the TNE value.

Action 2: Improved Experiment

The test data `sPrcn0915.csv` shows the results of our improved experimental for s*Prcn* characterization. The first row reflects the concentration of Co^{2+} (μM). The second row represents the intensity of fluorescence activity (i.e. Fluorescence/OD600). The third row is the concentration of IPTG (mM). In this experiment, a finer-scale sampling of Co^{2+} concentration is considered within the range of 0–100μM.

Turn to the "Trend" page and make a 3D Plot. This time we can find that: **(1) The intensity of fluorescence activity decreases with the increasing Co^{2+} concentration, which is still**

contradict to our expectation (we expected that low concentration of Co^{2+} should improve the expression of eGFP); (2) Different from the pre-experiment, the variation of Co^{2+}, rather than IPTG concentration, plays the major role in regulating the fluorescence activity. This conclusion can be further validated by GLM and ANOVA analyses.

The question (1) worth further discussion. Turn to the "Prediction" Page, we find that different interpolation methods can produce various results, which is puzzling. Interpolation methods that based on polynomial model (i.e. trend analysis and spline method) indicates that the expression of eGFP decreases with the increasing eGFP. For instance, Fig. S7 shows the results provided by spline method. However, methods based on local interpolation (i.e. inverse weighted interpolation, simple moving average, and kriging) shows a distinguish peak at low concentration of Co^{2+} (e.g. Fig. S8).

In uncertainty analysis, we can calculate the TNE value for each Co^{2+} concentration. The results show that the TNE value peaks at 25μM Co^{2+} (Fig. S9). Consequently, we still need to design finer-scale experiments. **Another quicker way is deleting all the observations with the treatment of 250μM or 500μM Co^{2+}**, because the TNE value at this region is closed to 0, indicating that observations located in this region is **"junk data"**.

Action 3: Delete Junk Data

After deleting the junk data, the 3D Plot exhibits excellent performance of fitting the data (Fig. S10). **The Fluorescence/OD600 value increases with the increasing of Co^{2+}. However, further decreasing is detected when the concentration of Co^{2+} exceeds 30μM.** This time, all interpolation methods give a similar result. In the end, Fig. S10 gives an ideal characterization of part sP*rcn*.

6.3 Case Study 2: Differential Expression of RNA

qRT-PCR is a laboratory technique of molecular biology based on RT-PCR. This technique monitors the amplification of targeted RNA molecular by inspecting its florescence intensity. Consequently, it is a semi-quantitative method which is helpful to quantify the expression level of mRNA. However, repetitive result is one of the biggest problems in qRT-PCR. Statistical analysis is thus critically important to ensure that your conclusion is robust.

This example shows how "Error" module in `Expmeasure` can help you to determine whether there is really a trend in your dataset. The test data is `repmiRNA.csv`.

Table 6.3.1. First three rows of the test data: `repmiRNA.csv`

Cell_Line	miRNA	Group	Relative	LOG
hek 293	22/U6	hek 293 22/U6	2.36e-05	-4.63
hek 293	195/U6	hek 293 195/U6	3.6e-04	-3.44
hek 293	199/U6	hek 293 199/U6	6.33e-06	-5.20

* "Cell_Line": type of cell; "miRNA": type of miRNA (U6 is the internal standard miRNA); "Group": a combination of "Cell_Line" and "miRNA"; "Relative": Relative expression level; "LOG": log transformation of relative expression level.

Different from the first case, the treatment in this dataset is categorical. In addition, we only need to test the difference of expression level between different cells, rather than quantifying a numerical trend. Consequently, violin plot is the best way to inspect the difference for different groups.

First, we perform a *t*-test for different miRNA using the following code in Rstudio:

```
dat <- read.csv(file.choose(), header=T) # read repmiRNA.csv
t22 <- t.test( dat[dat$Group=="hek293 22/U6",],
               dat[dat$Group=="hep G2 22/U6",] )
t195<- t.test( dat[dat$Group=="hek293 195/U6",],
               dat[dat$Group=="hep G2 195/U6",] )
t199<- t.test( dat[dat$Group=="hek293 199/U6",],
               dat[dat$Group=="hep G2 199/U6",] )
# You can also perform wilcox test using function `wilcox.test`
#     if you like
print(t22)
print(t195)
print(t199)
```

As you can see, the result is discouraging. No significant difference is found indicated by >0.1 p value. Is it means that there is really no trend in the dataset? Difference of two groups of data can be eliminate due to random error caused by equipment, especially for qRT-PCR. Thus, you can upload the data to `Expmeasure` and turn to "Error" page.

Before doing that, turn to "Trend" page and make a violin plot (Fig. S11). The result shows that the difference of miRNA expression between hek 293 and hep G2 does exist. Unfortunately, this trend may be reduced by some outliers (indicated by the wide range of boxes). Turn to the "Error" page, select both "Cell_Line" and "miRNA" to detect outliers. Delete the outliers in the source data file, then run the t-test script (mentioned above) and make a violin plot again. You will find the range of box becomes smaller. Again, you will find a outlier in your dataset when you perform algorithms in the "Error" page – "Outlier" panel and then delete outlier.

After repeating this for several times, you will find a distinctive difference between different cells in the violin plot (Fig. S12). Also, lower *p*-value will be found in the *t*-test.

7 Conclusion

All highlight points that `Expmeasure` (rather than any other software) gives you can be summarized as:

- "Uncertainty" module is helpful to **redesign your experiment**. It can remind you of potential strong trend hidden in your dataset due to incomplete experimental design. Also, TNE analysis can help you **identify "junk data"** (indicated by low TNE value). Deleting "junk data" is important to improve the model fitness.
- "Prediction" module provides a wide range of interpolation methods. **If different methods lead to different conclusions, be careful about potential finer-scale trend** in your dataset.
- "Error" module helps to identify outliers in your dataset. **Deleting outliers** is a sensible way to improve the model fitness, especially for categorical data.
- *K*-function is a statistic to **quantify first-order effects** hidden in the dataset. As an alternative method to test the significance of any trend in your dataset, *K*-function analysis makes your conclusion more robust.

Enjoy your journey in `Expmeasure` and hope that you can get a sensible result with more confidence.

Appendix 1. Install `Expmeasure`

As described in Section 1.2, github is the best way to install `Expmeasure`. However, if you fail to install `Expmeasure` with unexpected error message, there is another way to run this APP.

You will find a Rscript version of `Expmeasure` in the following url:

https://2021.igem.org/Team:ZJU-China/Software

Download the Rscript `expmeasure.R`, and open the Rstudio software. Click "source" button to load the `expmeasure` function. Load the dependence packages before running the APP:

```
library(shiny)
library(data.table)
library(shinydashboard)
library(shinyWidgets)
library(plotly)
library(agricolae)
library(tidyverse)
library(ggplot2)
library(rgl)
library(ggforce)
library(ggpubr)
library(car)
library(sp)
library(gstat)
library(raster)
library(fields)
library(bootstrap)
library(shinythemes)
```

If you don't have these packages, install them from CRAN using the command `install.packages("xxx")`. After that, you can run the APP using the command `expmeasure()`. You will see the HOME page of our software in the browser that is launched by `Expmeasure` (Fig. 1.3.1).

Appendix 2. Supplementary Figures

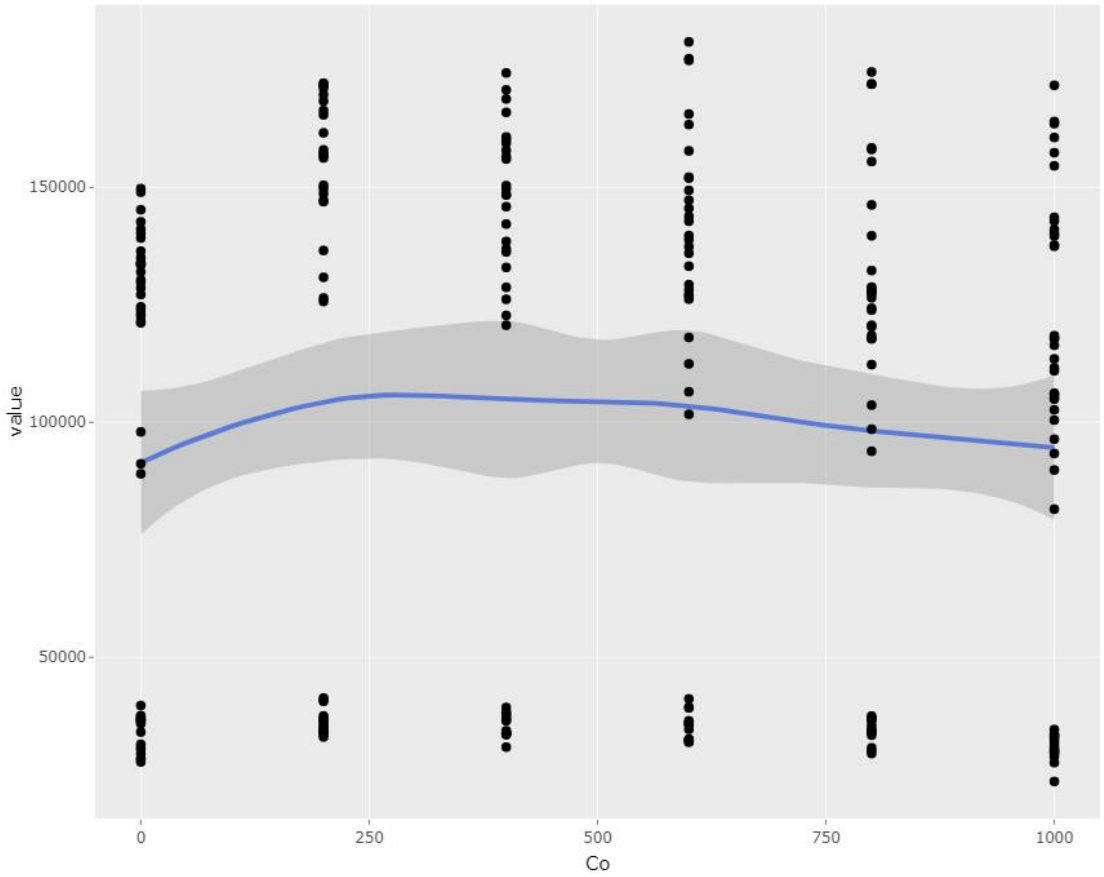

Figure S1. Relationship between the intensity of fluorescence activity and concentration of Co^{2+} in the pre-experiment of sP*rcn* characterization.

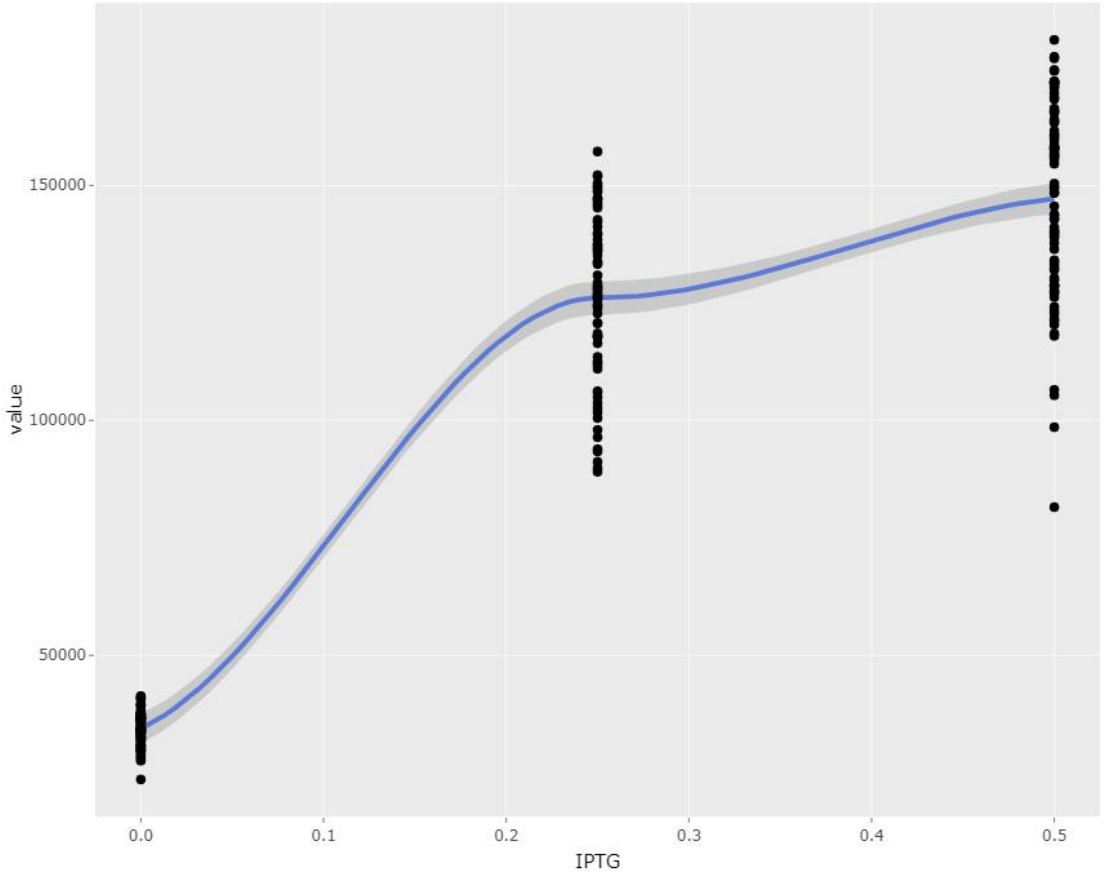

Figure S2. Relationship between the intensity of fluorescence activity and concentration of IPTG in the pre-experiment of sP*rcn* characterization.

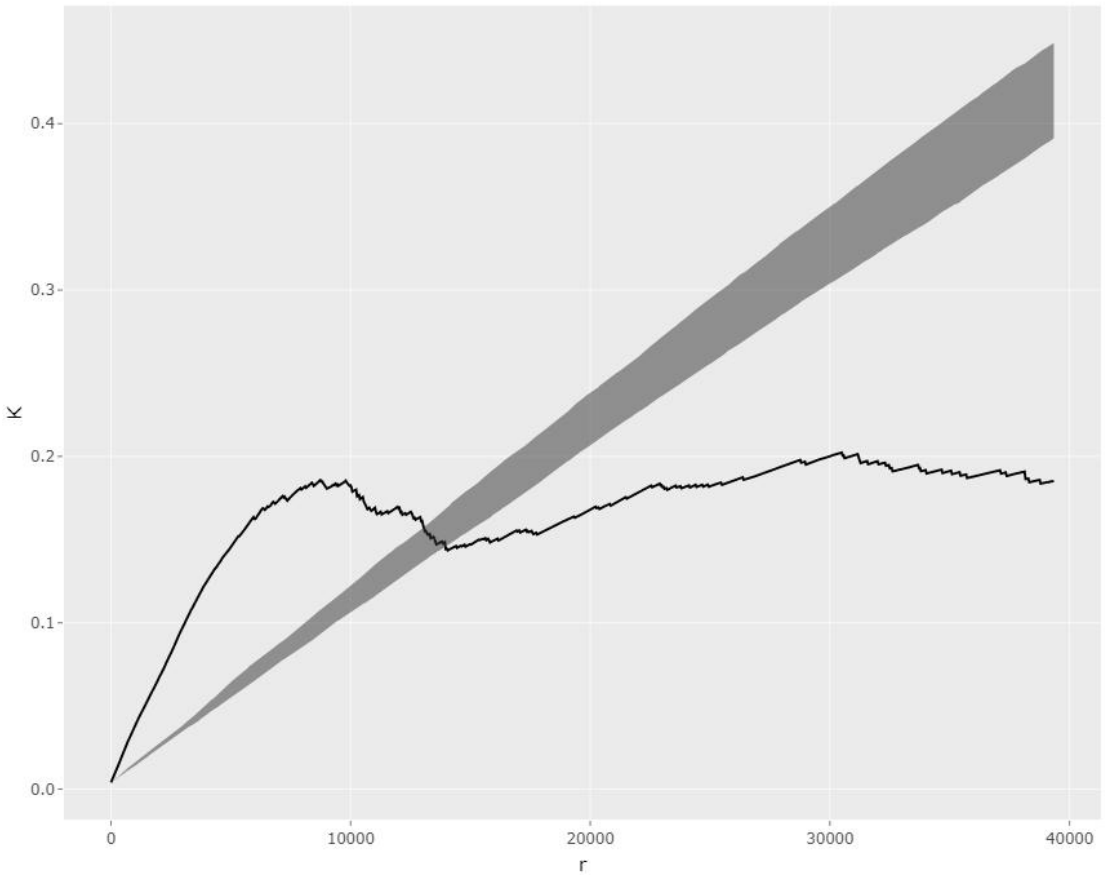

Figure S3. *K*-function plot of Co^{2+} concentration

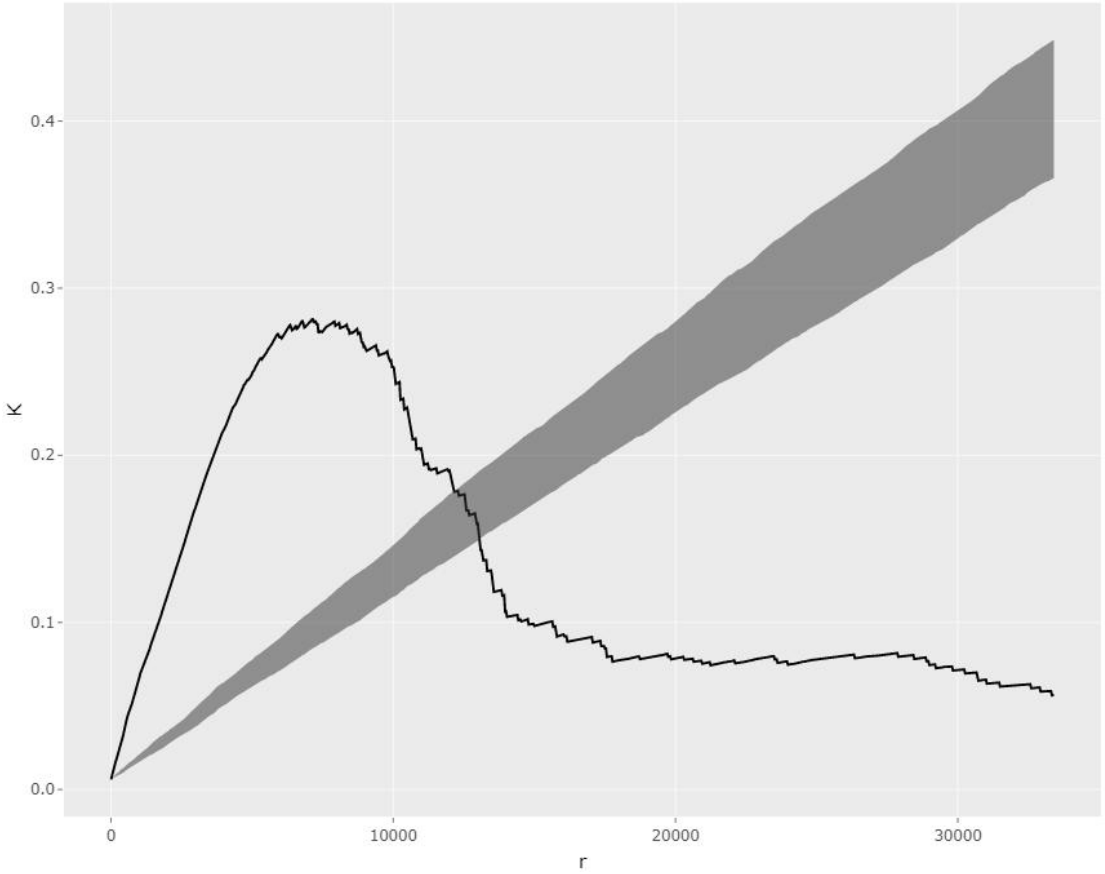

Figure S4. *K*-function plot of IPTG concentration

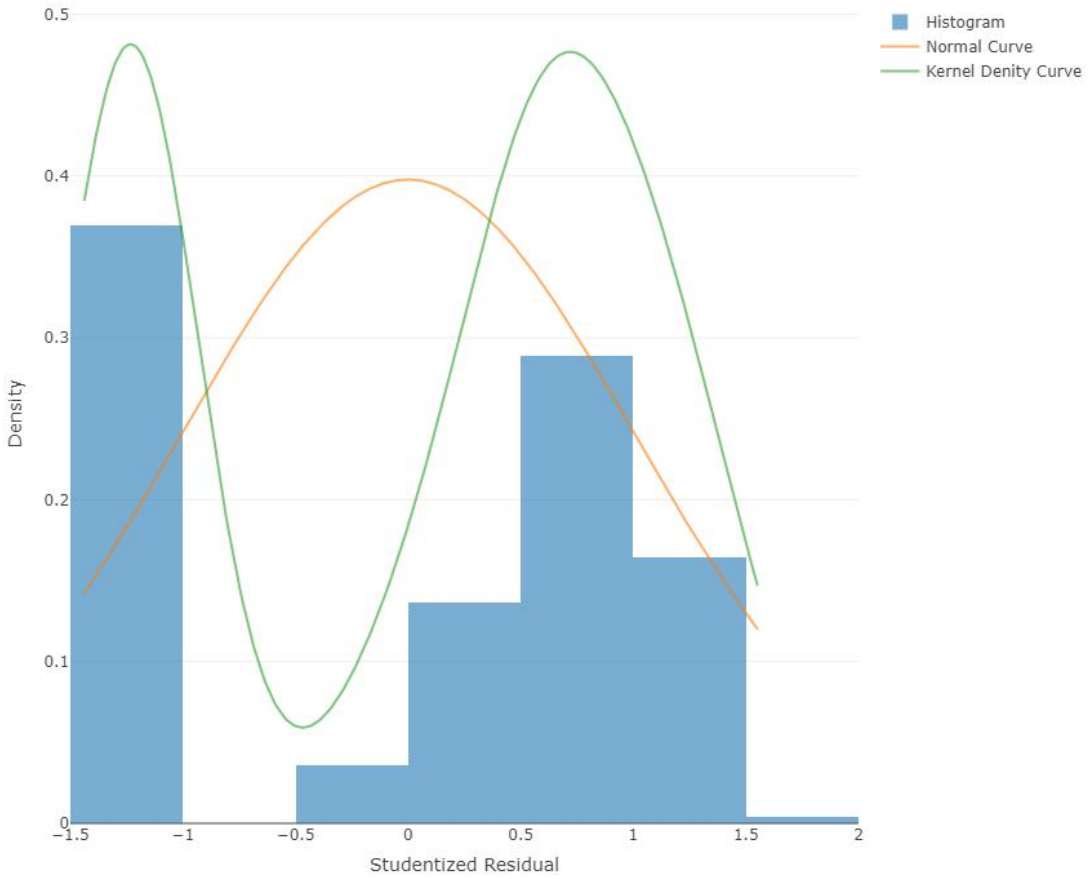

Figure S5. The residue plot when GLMs only contains one explanatory variable – the concentration of Co^{2+}

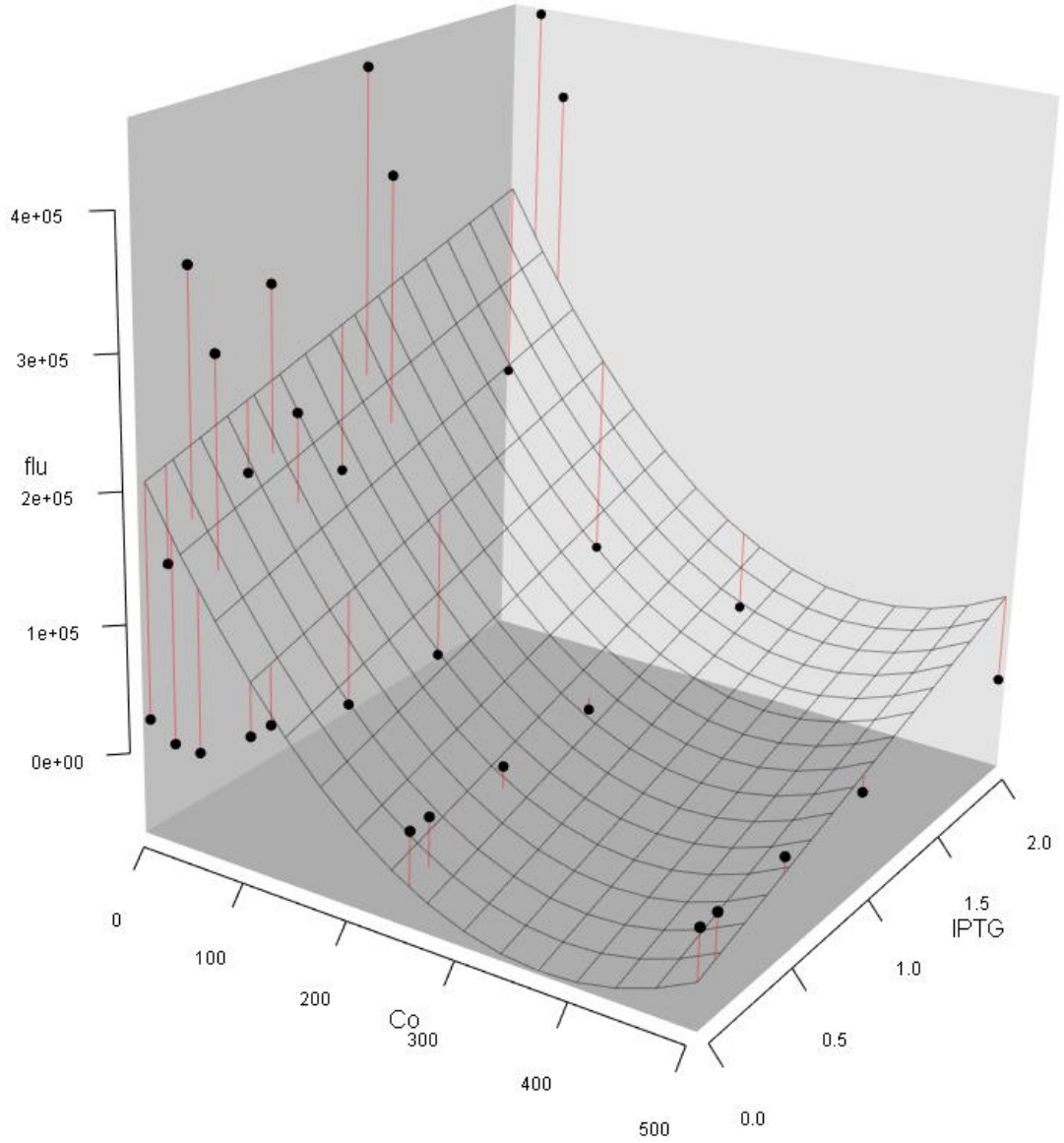

Figure S6. 3D Plot that reflects the relationship among Co^{2+}, IPTG concentration, and the intensity of fluorescence activity in the improved experiment for sP*rcn* characterization.

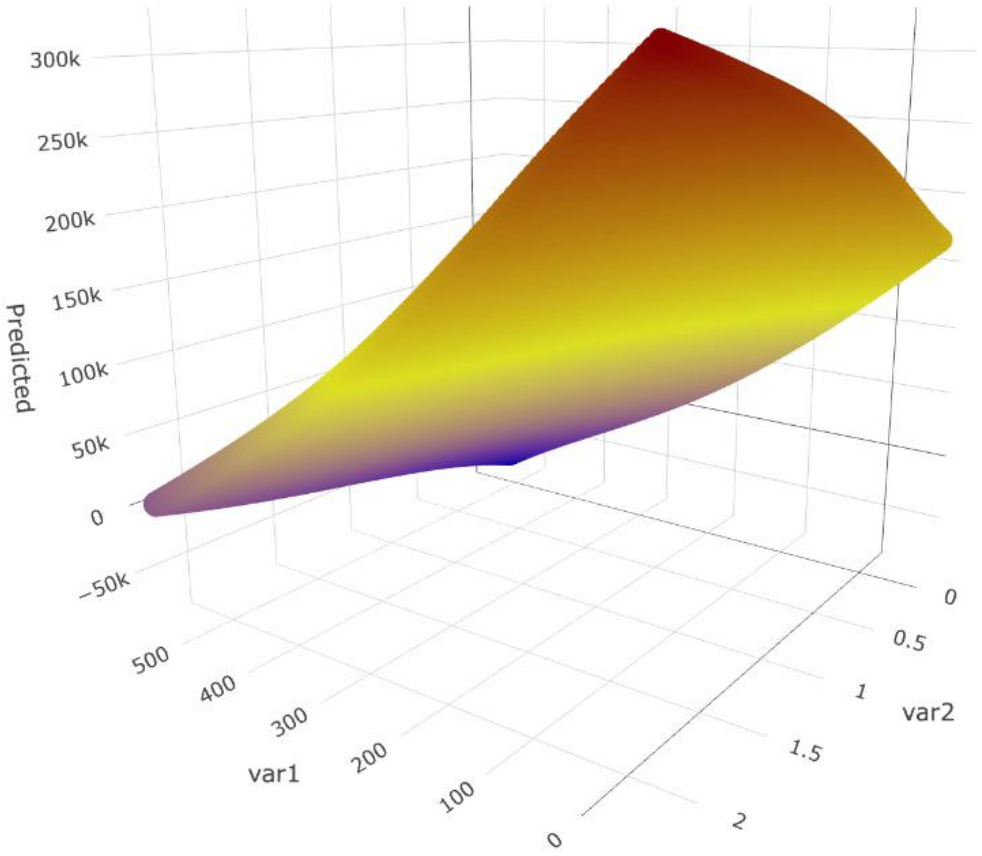

Figure S7. Spline interpolation of the dataset `sPrcn0915.csv`. var1 represents the concentration of Co^{2+}, while var2 represents the concentration of IPTG. The *z*-axis represents the intensity of fluorescence activity.

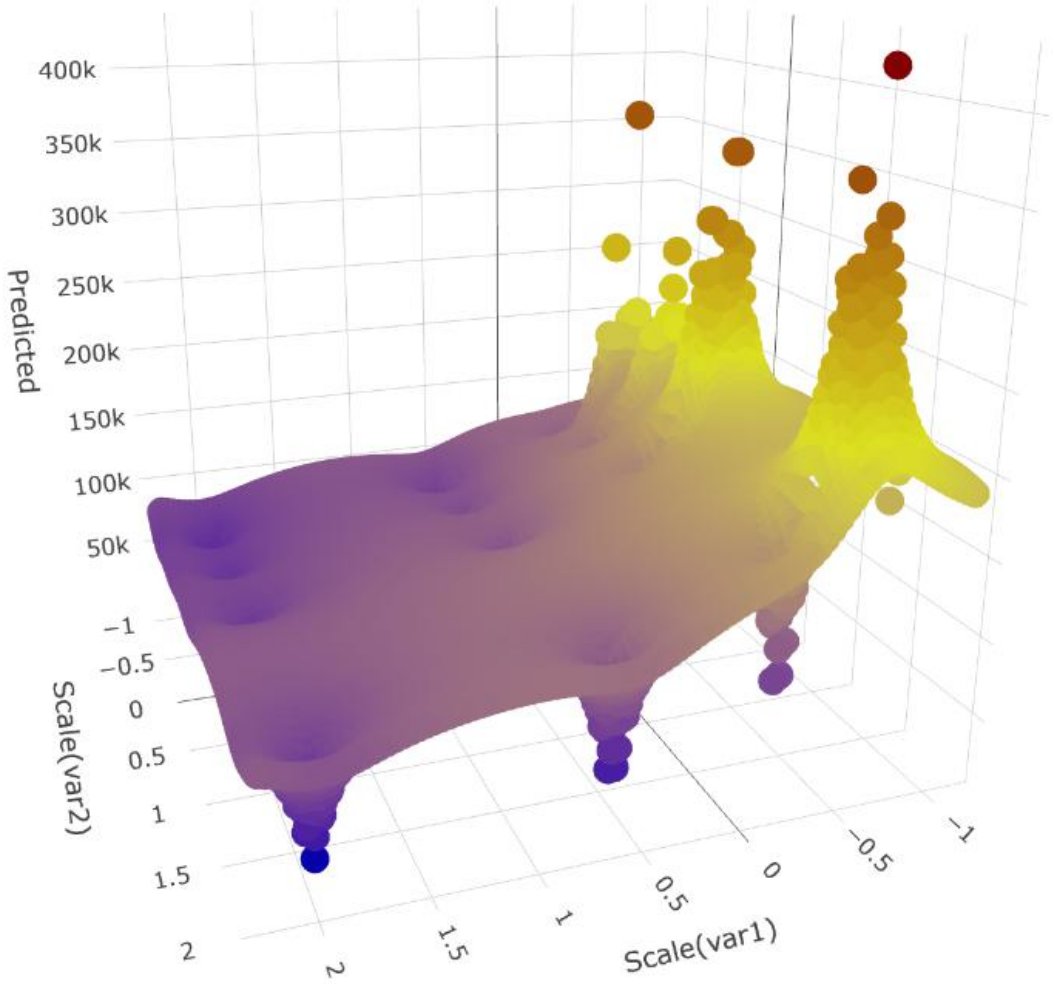

Figure S8. Inverse weighted interpolation of the dataset `sPrcn0915.csv`. var1 represents the concentration of Co^{2+}, while var2 represents the concentration of IPTG. The *z*-axis represents the intensity of fluorescence activity.

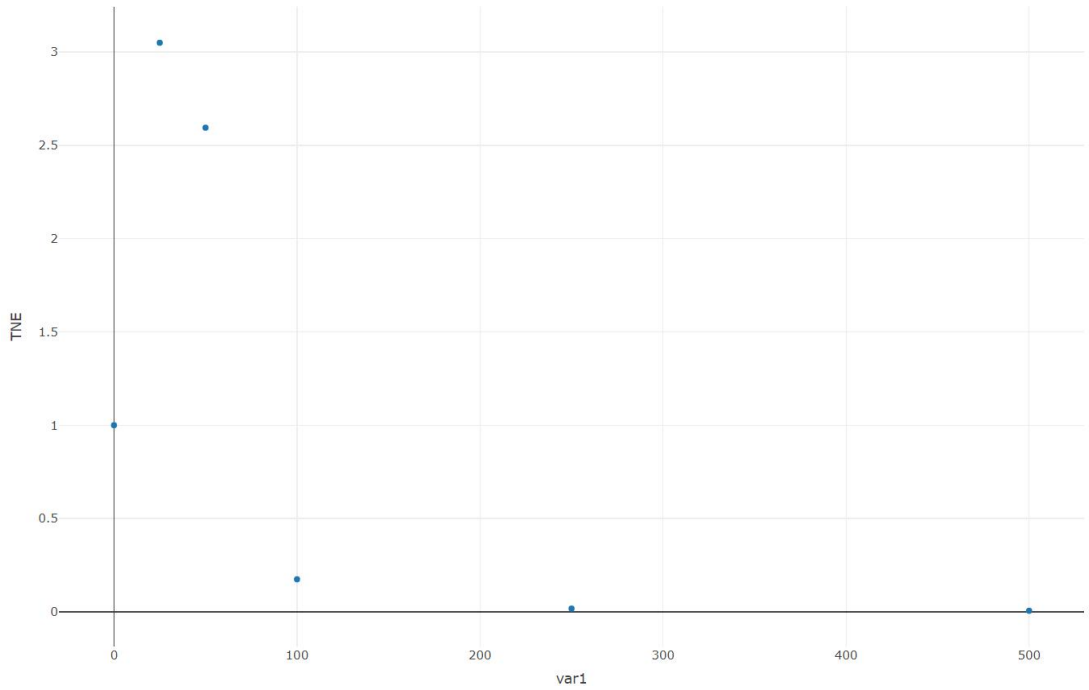

Figure S9. The TNE plot of the dataset `sPrcn0915.csv`. var1 represents the concentration of Co^{2+}.

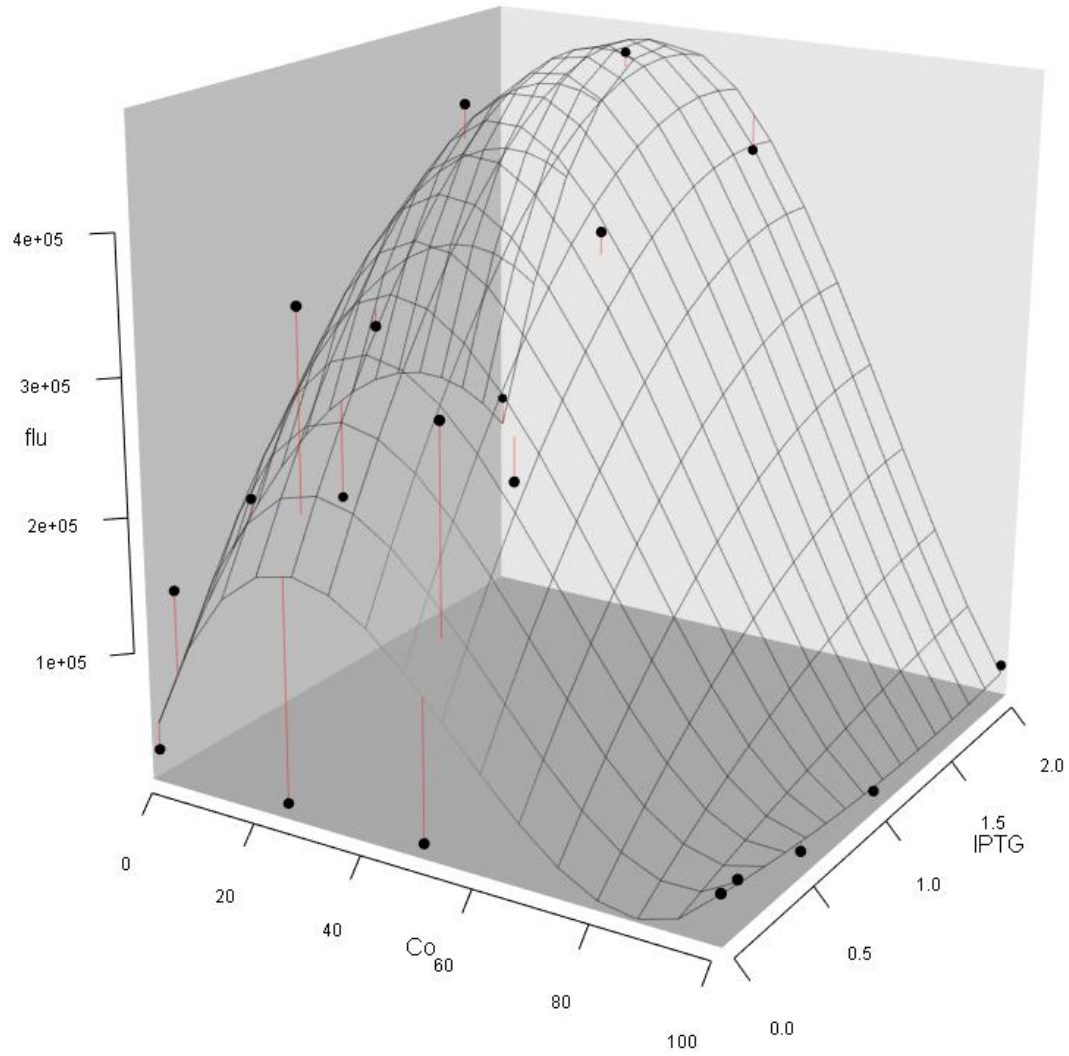

Figure S10. 3D Plot that reflects the relationship among Co^{2+}, IPTG concentration, and the intensity of fluorescence activity in the improved experiment for s*Prcn* characterization after deleting "junk data".

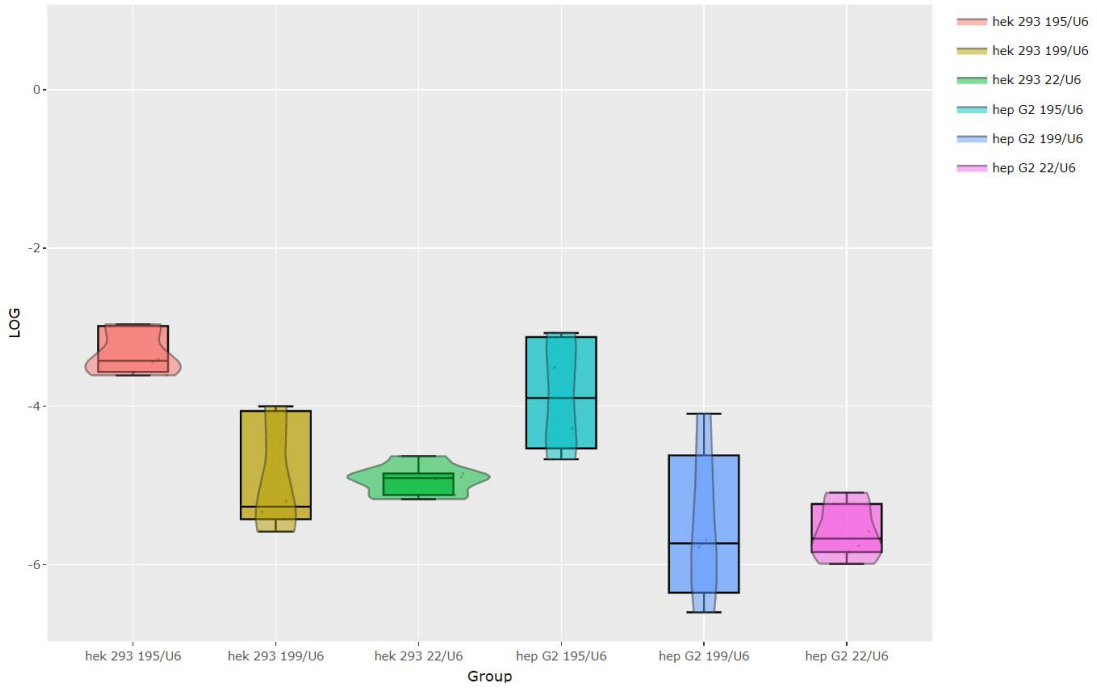

Figure S11. The violin plot of miRNA expression levels in different cells. The y-axis indicates the relative expression level of corresponding miRNA (either 195, 199, or 22).

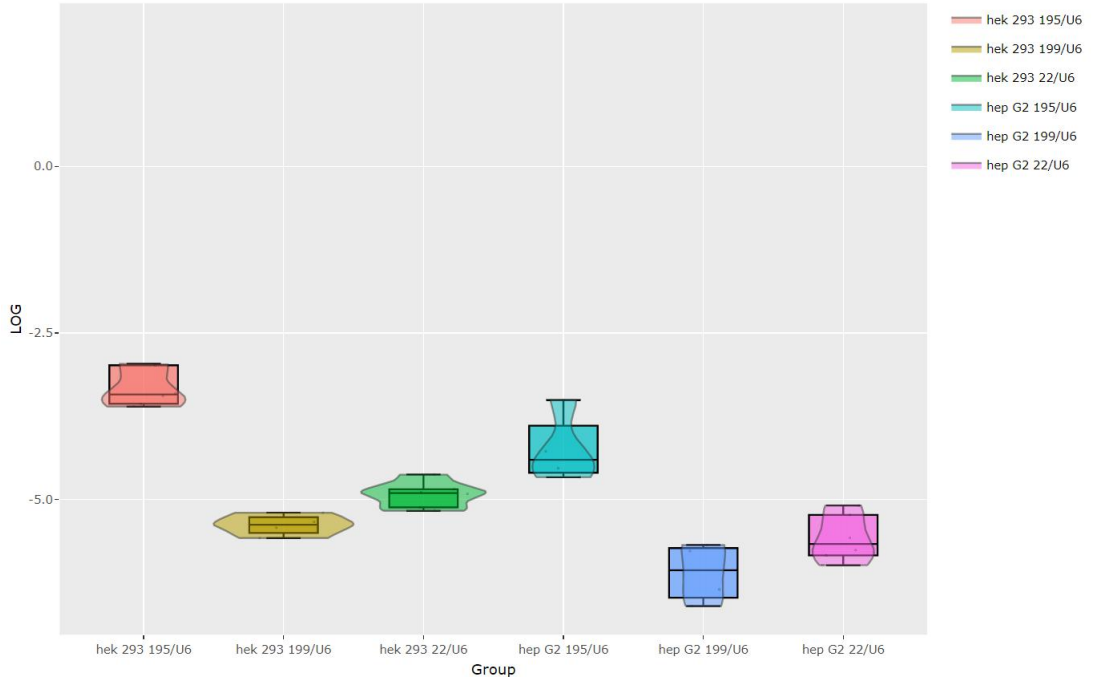

Figure S11. The violin plot of miRNA expression levels in different cells (After deleting the outliers). The y-axis indicates the relative expression level of corresponding miRNA (either 195, 199, or 22).