Final Project

Phase 2 – Progress Report

Lee Panter

Preface

Hi Matt and/or Harris!

Sorry for the length of this submission. I promise it was not my intention to submit something so long, but I did want to cover all of the Phase Two objectives. There is probably A LOT that you can ignore in here, and probably some code (and code output) that you probably did not want to see. I'm very sorry to waste your time if this is the case. If you want me to cut this document down, I am happy to try, but I feel like the entire process follows the "Description" fairly strictly.

Please let me know if this submission is unacceptable. Thank you.

Description

This script/writeup entails Phase Two of the final project in BIOS 6643. It contains:

- Details about the finalized data, including:
  - Variable definitions, and nomeclature
  - Explanatory layouts
  - Reduction measures for introductory analyses
  - Exploratory data analyses (both graphical and numerical)
    - Exploratory data analyses of "log-transformed" variables
- Initial Modeling Methodologies, including:
  - Theoretical modeling formats
  - Essential modeling assumptions (outcome/predictor types)
- Basic results from model fits
  - Streamlined code outputs
  - Small paragraph of explanation
  - Probable next steps

Final Data Information

Basic Data Format

The data on which the following analyses are conducted can be best represented in "short format" after the definition of a couple noteable variable sets, and indices.

Nomeclature and Variables

We define the set of RNAseq (RNA sequencing) variables using:

$$\begin{aligned}\Omega_{SEQ} &= \{SEQ_p : \ p = 1, \ldots, 38354\} \\ &= \{SEQ_1 = \text{“}A1BG - AS1\text{”}, \ SEQ_2 = \text{“}A1CF\text{”}, \ \ldots, \ SEQ_{38353} = \text{“}SNOZ5\text{”}, \ SEQ_{38354} = \text{“}SNOSNR66\text{”}\}\end{aligned}$$

In this manner, we may refference $SEQ_{3350} =$ "$AL445384.1$". Note that $|\Omega_{SEQ}| = 38354$

We also define the set of Flow Cytometry variables using:

$$\begin{aligned}\Omega_{FLOW} &= \{FLOW_p : \ p = 1, \ldots, 19\} \\ &= \{FLOW_1 = \text{“}FSC.A\text{”}, \ FLOW_2 = \text{“}FSC.W\text{”}, \ \ldots, \ FLOW_{18} = \text{“}CD27\text{”}, \ FLOW_{19} = \text{“}CD235a\text{”}\}\end{aligned}$$

In this manner, we may refference $FLOW_{14} =$ "$CD31$". Note that $|\Omega_{FLOW}| = 19$

and we define the set of MetaData variables using:

$$\begin{aligned}\Omega_{META} &= \{META_p : \ p = 1, \ldots, 14\} \\ &= \{META_1 = \text{“}measurement.name\text{”}, \ \ldots, \ META_{14} = \text{“}Perc.Mt\text{”}\}\end{aligned}$$

In this manner, we may refference $META_8 =$ "$CD31$". Note that $|\Omega_{META}| = 14$

following with tradition we will be using $i = 1, \ldots, 15$ to symbolize subject, and $j = 1, \ldots, n_i$ to symbolize the number of repeated measurements within subject i.

With these definitions the short-versions of the data become much more illuminating.

RNAseq Data

The short-format data for the RNAseq data may be written as:

$i \downarrow \ p \rightarrow$	$(SEQ_1)_{ij}$	$\cdots$	$(SEQ_{38354})_{ij}$
i=1	$(SEQ_1)_{11}$	$\cdots$	$(SEQ_{38354})_{11}$
	$\vdots$	$\vdots$	$\vdots$
i=1	$(SEQ_1)_{1n_1}$	$\cdots$	$(SEQ_{38354})_{1n_1}$
$\vdots$			
i=15	$(SEQ_1)_{15\ 1}$	$\cdots$	$(SEQ_{38354})_{15\ 1}$
	$\vdots$	$\vdots$	$\vdots$
i=15	$(SEQ_1)_{1\ n_{15}}$	$\cdots$	$(SEQ_{38354})_{1\ n_{15}}$

Flow Cytometry Data

The short-format data for the Flow Cytometry data may be written as:

$i \downarrow \ p \rightarrow$	$(FLOW_1)_{ij}$	$\cdots$	$(FLOW_{19})_{ij}$
i=1	$(FLOW_1)_{11}$	$\cdots$	$(FLOW_{19})_{11}$
	$\vdots$	$\vdots$	$\vdots$
i=1	$(FLOW_1)_{1n_1}$	$\cdots$	$(FLOW_{19})_{1n_1}$
$\vdots$			
i=15	$(FLOW_1)_{15\ 1}$	$\cdots$	$(FLOW_{19})_{15\ 1}$
	$\vdots$	$\vdots$	$\vdots$
i=15	$(FLOW_1)_{1\ n_{15}}$	$\cdots$	$(FLOW_{19})_{1\ n_{15}}$

Reduction Measures for Introductory Analyses

For the purpose of introductory analyses the full data is reduced to a single predictor and single outcome variable for each of the Flow and RNAseq data frames. Specifically, we perform initial regression anlyses on the RNAseq variables: $SEQ_{35858} =$ "$TNKS$", and $SEQ_{6202} =$ "$CD22$". (a univariate Flow analysis will follow if time allows).

Quality Control Measure for Final Data

It should be noted that the data used here, and for further analyses has previously passed Quality Control measures imposed from a combination of sources:

- The data source (original research publication) (Arazi et al. 2018)
  - All qualifying cells must have between 1,000 and 5,000 detected genes (1,000 < nFeature < 5,000)
- The Seurat tutorial for Quality Control ("Satija Lab," n.d.)
  - All qualifying cells must be filtered according to a maximum percentage threshold of Mitochondrial DNA (threshold altered from recommended value specified)
- Outside Industry Sources
  - Setting DNA threshold to 60% is appropriate
  - Subsetting to B-Cells only

Exploratory Data Analysis

Quantitative Data Summaries

We produce six number summaries of the data variables $TNKS$ and $CD22$

TNKS

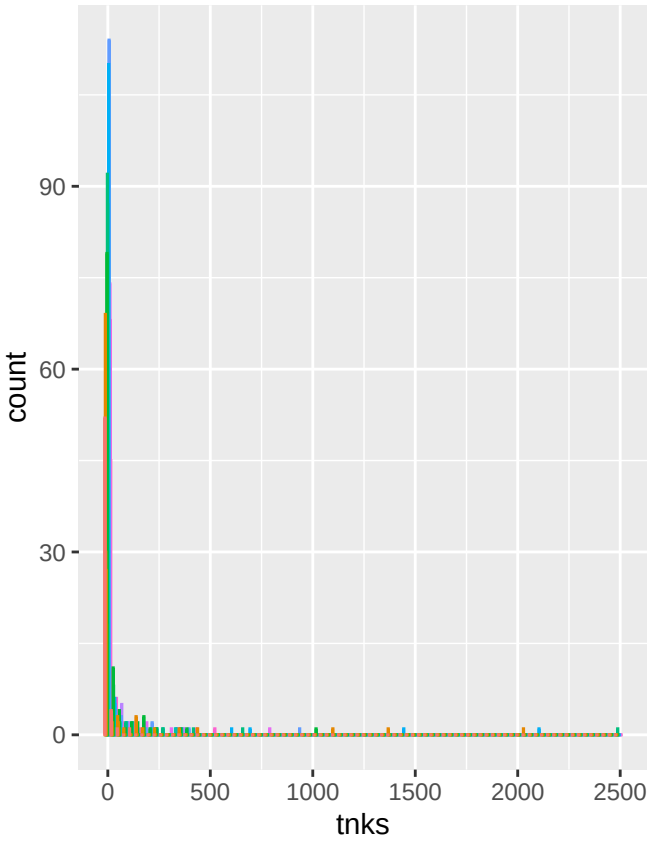
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   29.73    3.00 2493.00
```

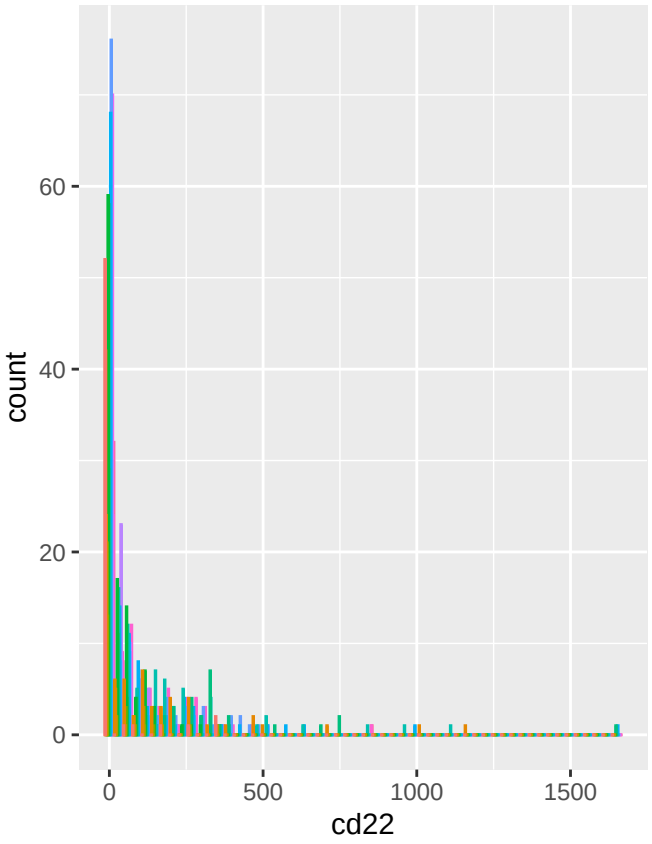
CD22

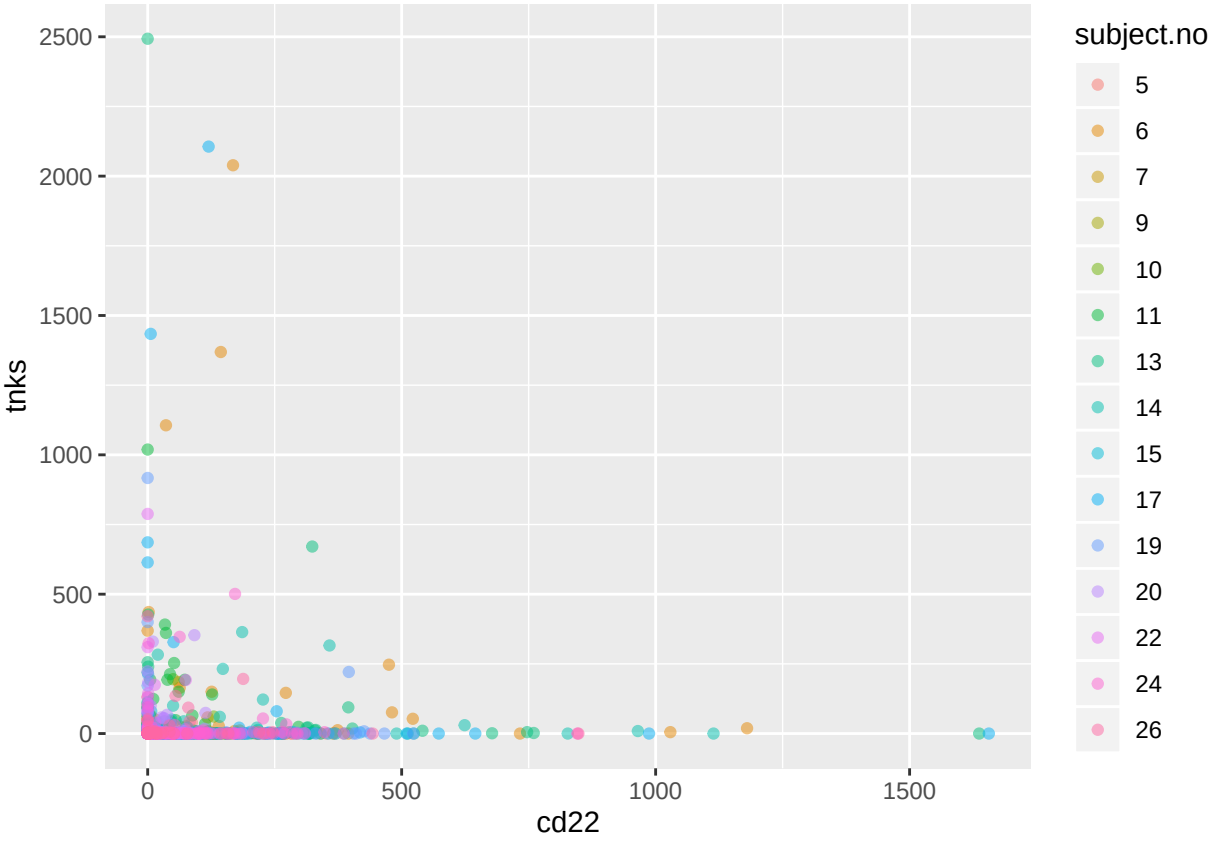
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    6.50   70.22   73.00 1656.00
```

Graphical Data Summaries

We produce histograms of $TNKS$ and $CD22$, and a scatter plot of their relationship.







Log-Transformed Variables

Given the severity of the right skew in the variables displayed above, it would be wise to also perform a log-transformation as a prepatory action. We also display these transformed variables:

We produce six number summaries of the data variables $log(TNKS + 1)$ and $log(CD22 + 1)\backslash$

log(TNKS+1)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.9796  1.3863  7.8216
```

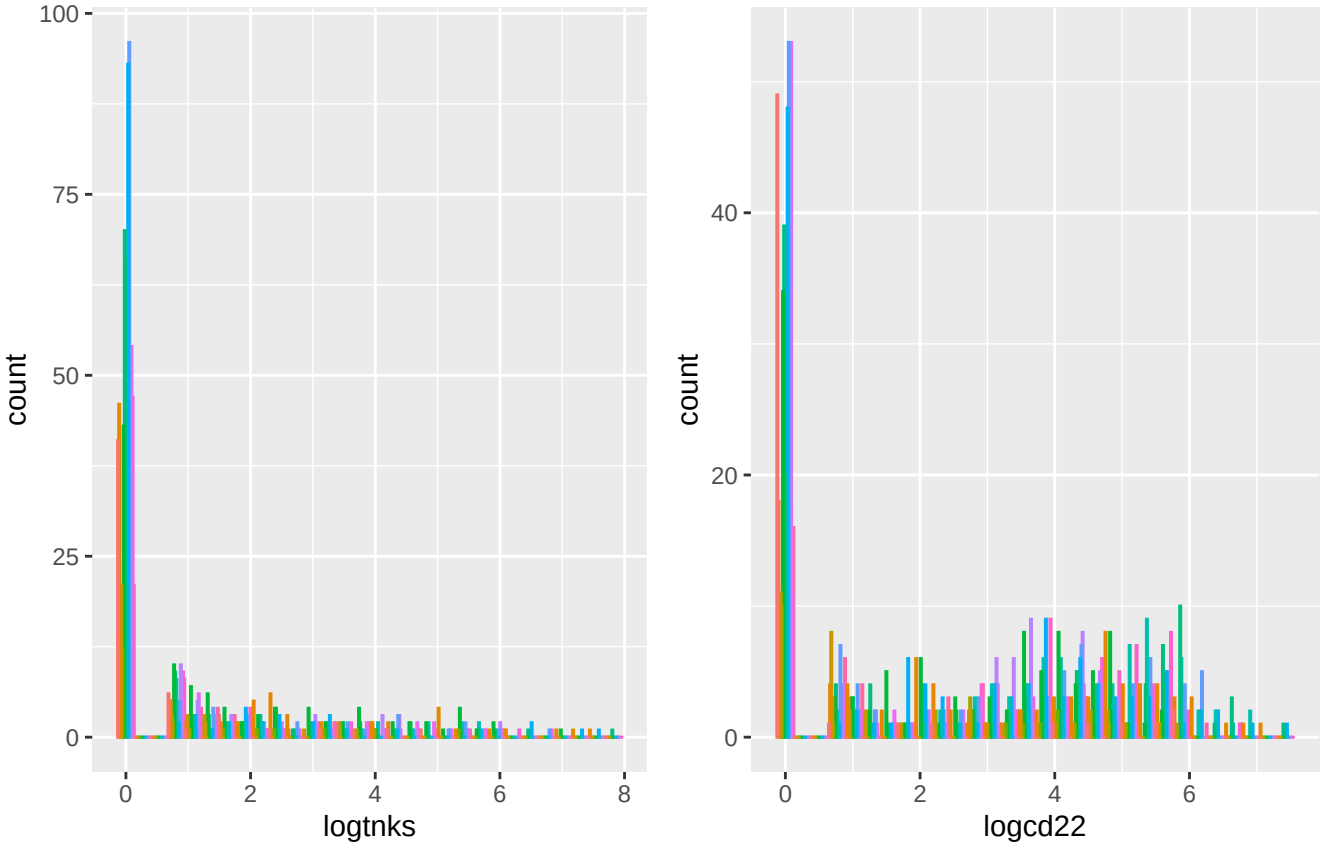
log(CD22+1)

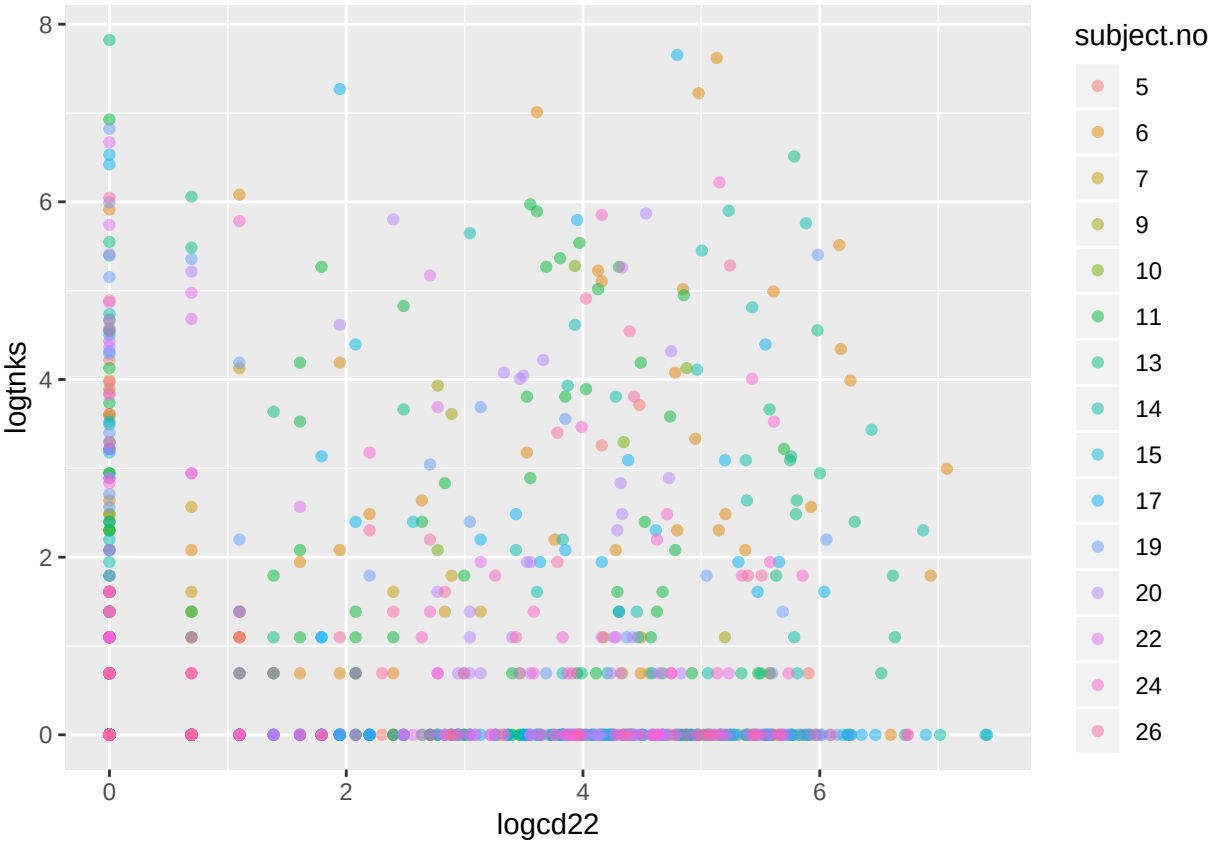
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.013   2.311   4.304   7.413
```

Graphical Data Summaries

We produce histograms of $log(TNKS + 1)$ and $log(CD22 + 1)$, and a scatter plot of their relationship.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```





Initial Modeling Methodologies

Due to the intrinsic "count" nature dictated by the ourcome variable of interest, the ultimate goal of this project is to compare the effect of models that meet experimental design at different levels.

Specifically, we look to take three approaches to modeling the Seqencing Variable $TNKS$:

1. Brute Force: Linear Model-OLS, and Linear Mixed Effects (least-good fits)
2. Some Finesse: Poisson Regression-no over-dispersion and quasilikelikelihood (better fits), also incorperating Random Effect methodologies if time permits
3. Better Still: Zero-Inflated Poisson and Zero-inflated Negative Binomial (even better fit)

In all cases, it will be assumed that responses are independent between subjects AND within subjects. This can be reasonably justified in the context of the experimental design considering the sampling technique.

Please note that indices have been kept general in the notation below to allow for future, more general models that encompass more variables (should time allow). In the notation that follows we have

$$\left(Y_{p_{resp}^{seq}}\right)_{ij} = TNKS_{ij}$$

and

$$\left(X_{p_{cov}^{seq}}\right)_{ij} = CD22_{ij}$$

and models that incorperate the "flow" subscript are incorperated to allow for modeling Flow Cytometry data in a similar fashion.

Brute Force Modeling Information

These models are NOT an attempt to optimize fit, accuracy or any metric. They are going to be used for baseline comparison.

The following mixed (incorperating both fixed and random effects will be evaluated)

Model 1: Linear Model with Fixed Effects (LMwFE)

$$\left(Y_{p_{resp}^{seq}}\right)_{ij} = \beta_0 + \beta_{1i} + \beta_{2i}\left(X_{p_{cov}^{seq}}\right)_{ij} + \epsilon_{ij}$$

and

$$\left(Y_{p_{resp}^{flow}}\right)_{ij} = \beta_0 + \beta_{1i} + \beta_{2i}\left(X_{p_{cov}^{flow}}\right)_{ij} + \epsilon_{ij}$$

where

$$\epsilon_{ij} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}\right)$$

Model 2: Linear Mixed Model with Random Effects (LMMwRE)

$$\left(Y_{p_{resp}^{seq}}\right)_{ij} = \beta_0 + \beta_{2i}\left(X_{p_{cov}^{seq}}\right)_{ij} + b_{0i} + b_{1i}\left(X_{p_{cov}^{seq}}\right)_{ij} + \epsilon_{ij}$$

and

$$\left(Y_{p_{resp}^{flow}}\right)_{ij} = \beta_0 + \beta_{2i}\left(X_{p_{cov}^{flow}}\right)_{ij} + b_{0i} + b_{1i}\left(X_{p_{cov}^{flow}}\right)_{ij} + \epsilon_{ij}$$

where

$$\mathbf{b} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(\mathbf{0}, \mathbf{G})$$

$$\mathbf{G} = \begin{bmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

$$\epsilon_{ij} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}\right)$$

Minor Finnese Approaches

We hypothesize these approaches to be an improvement over the "Brute Force" approaches, but do not expect these models to encapsulate the full extent of the data's mean or variational behavior.

We intend to model the following Poisson Regression models:

Model 3: Poisson Regression without over-dispersion (POI)

The three elements of the GzLM are given by:

1. We assume that $TNKS \sim Poisson\,(\lambda)$ (Responses are distributed according to the Poisson exponential family)
2. We use will use the linear predictor:
$$\eta = \beta_0 + \beta_1\left(X_{p_{cov}^{flow}}\right)$$
3. The linear predictor (η) and the mean response ($\mu = E\,[Y|X]$) are related by the link function $g(x) = \log(x)$ according to the rule:
$$\mu = g^{-1}\,(\eta)$$

Model 4: Quasi-likelihood Poisson Regression (POIql)

The three elements of the GzLM are given by:

1. We assume that $TNKS \approx Poisson(\lambda)$ with $\sigma^2 = var\,(TNKS) = \phi\lambda$, i.e responses are "distributed" according to the quasi-distribiution Poisson eponential family.
2. We use the linear predictor:
$$\eta = \beta_0 + \beta_1 \left(X_{p_{cov}^{flow}}\right)$$
3. The linear predictor (η) and the mean response ($\mu = E\,[Y|X]$) are related by the link function $g(x) = \log(x)$ according to the rule:
$$\mu = g^{-1}\,(\eta)$$

Better-Still Models

Our most optimally suited models under consideration are a class of Mixture Models called "Zero-Inflated" models. These models are based on an underlying mixture of zeros and (in this case) a positive count outcome.

Model 5: Zero-Inflated Poisson (ZIP)

This may be thought of as:

> a regular random-event process, taking place in unit time, containing excess zero-count(s) ("Zero-Inflated Model - Wikipedia," n.d.)

and we may write this model as a mixture:

$$\mathbf{P}\,(y_{ij} = k) = \begin{cases} \mathbf{P}\,(y_{ij} = 0) = p \\ \mathbf{P}\,(y_{ij} = k) = (1-p)\,\frac{\lambda^k e^{-\lambda}}{k!} \end{cases}$$
$$\forall k \in \mathbb{N}$$

Model 6: Zero-Inflated Negative Binomial (ZINB)

This is simply an alteration of model 5, and will hypothetically only show neglible improvements unless the underlying distribution of non-zero data is distinctly non-poisson.

$$\mathbf{P}\,(y_{ij} = k) = \begin{cases} \mathbf{P}\,(y_{ij} = 0) = p \\ \mathbf{P}\,(y_{ij} = k) = (1-p)\begin{pmatrix} k+r-1 \\ k \end{pmatrix}(1-\lambda)^r\,\lambda^k \end{cases}$$

Basic Results from Model Fits

At this time, we (I) have only started to work on linear and linear mixed models. Below you will see the code I have used to create these two models, some diagnostic tables (R summary tables), and three diagnostic plots of each model:

- Model Plotted on Original Data
- Residual vs Fitted Values
- QQ Plots

In each I will also describe if there are:

1. Outstanding issues in model fit that need to be addressed
2. Further considerations for model comparisons
3. Next Steps (if applicable)

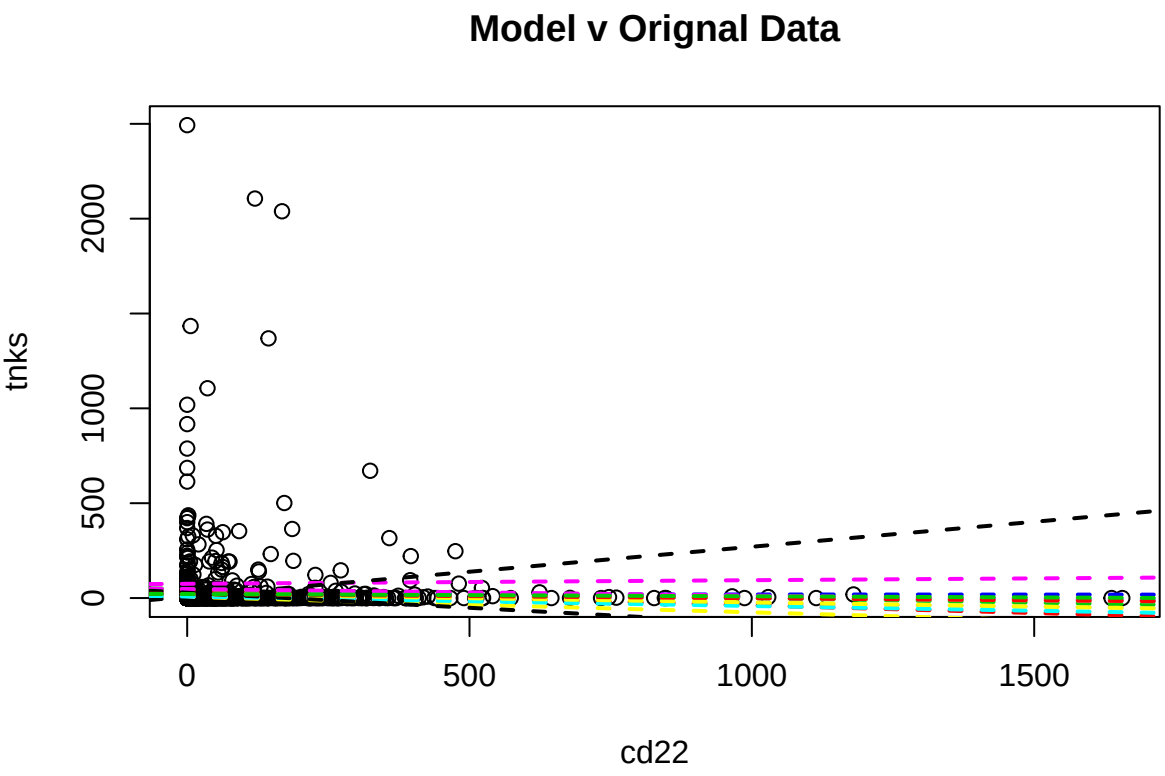
Model 1: Linear Model with Fixed Effect (LMwFE)

```
LMwFE=lm(tnks~subject.no + subject.no:cd22, data=dat)
LMwFEs=summary(LMwFE)
LMwFEs
```

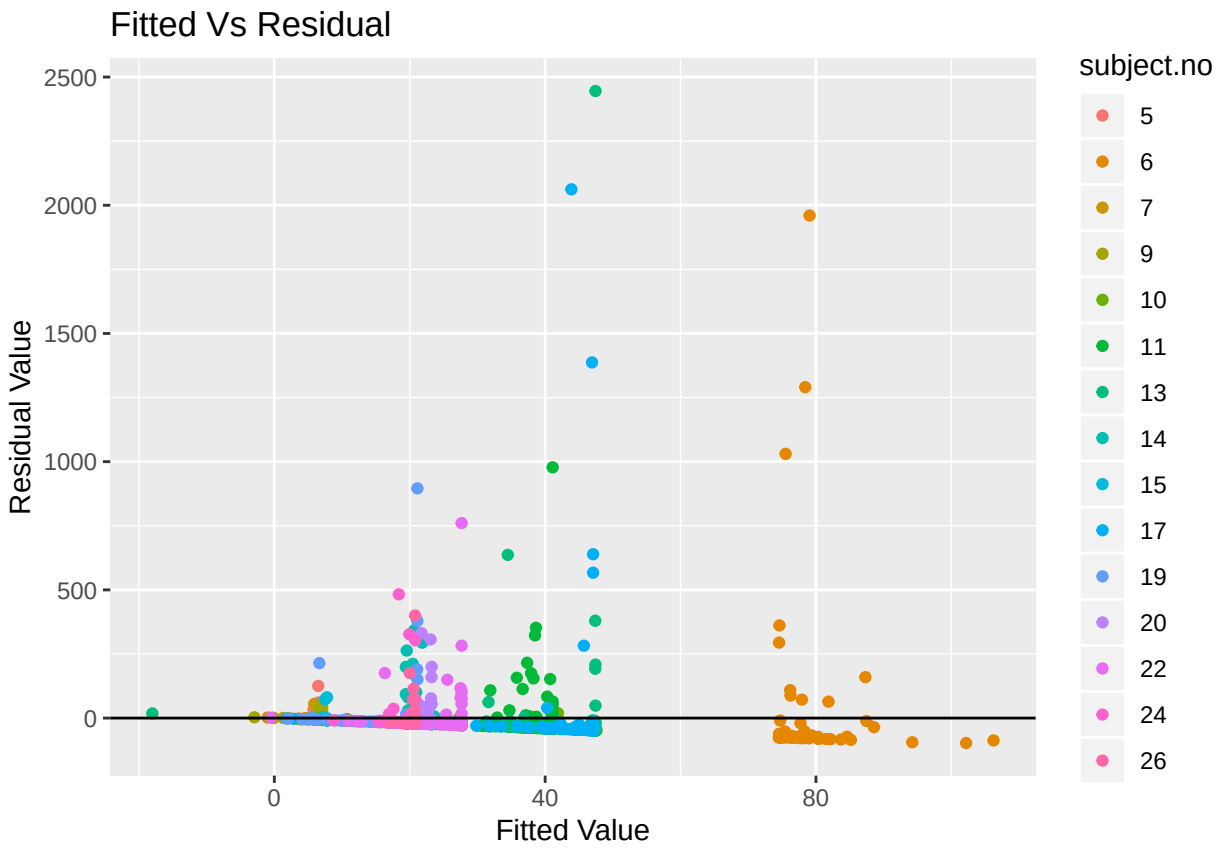
```
##
## Call:
## lm(formula = tnks ~ subject.no + subject.no:cd22, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -97.18  -38.81  -20.77   -7.79 2445.56
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.488246  21.311653   0.304   0.7608
## subject.no6        68.068161  28.766407   2.366   0.0181 *
## subject.no7        -0.450031  37.526737  -0.012   0.9904
## subject.no9         0.578969  38.215554   0.015   0.9879
## subject.no10        0.174449  46.286927   0.004   0.9970
## subject.no11       34.623121  27.893165   1.241   0.2148
## subject.no13       40.950295  27.182724   1.506   0.1322
## subject.no14       12.938384  28.456144   0.455   0.6494
## subject.no15        1.297663  38.772424   0.033   0.9733
## subject.no17       40.598442  26.165610   1.552   0.1211
## subject.no19       14.648172  26.602493   0.551   0.5820
## subject.no20       16.754000  31.822824   0.526   0.5987
## subject.no22       21.179817  28.082325   0.754   0.4509
## subject.no24       14.357018  30.866854   0.465   0.6419
## subject.no26       14.301006  31.512122   0.454   0.6500
## subject.no5:cd22   -0.007811   0.267016  -0.029   0.9767
## subject.no6:cd22    0.026846   0.080754   0.332   0.7396
## subject.no7:cd22   -0.073982   1.139637  -0.065   0.9483
## subject.no9:cd22   -0.027427   0.326802  -0.084   0.9331
## subject.no10:cd22   0.271102   0.784186   0.346   0.7296
## subject.no11:cd22  -0.072486   0.266263  -0.272   0.7855
## subject.no13:cd22  -0.039984   0.068080  -0.587   0.5571
## subject.no14:cd22   0.006764   0.078073   0.087   0.9310
## subject.no15:cd22  -0.042184   1.098959  -0.038   0.9694
## subject.no17:cd22  -0.026714   0.069803  -0.383   0.7020
## subject.no19:cd22  -0.036552   0.112542  -0.325   0.7454
## subject.no20:cd22  -0.016174   0.431567  -0.037   0.9701
## subject.no22:cd22  -0.151140   0.505629  -0.299   0.7651
## subject.no24:cd22  -0.014250   0.133672  -0.107   0.9151
## subject.no26:cd22  -0.004135   0.154920  -0.027   0.9787
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 155.6 on 1080 degrees of freedom
## Multiple R-squared:  0.01534,    Adjusted R-squared:  -0.0111
## F-statistic:  0.58 on 29 and 1080 DF,  p-value: 0.9637
```

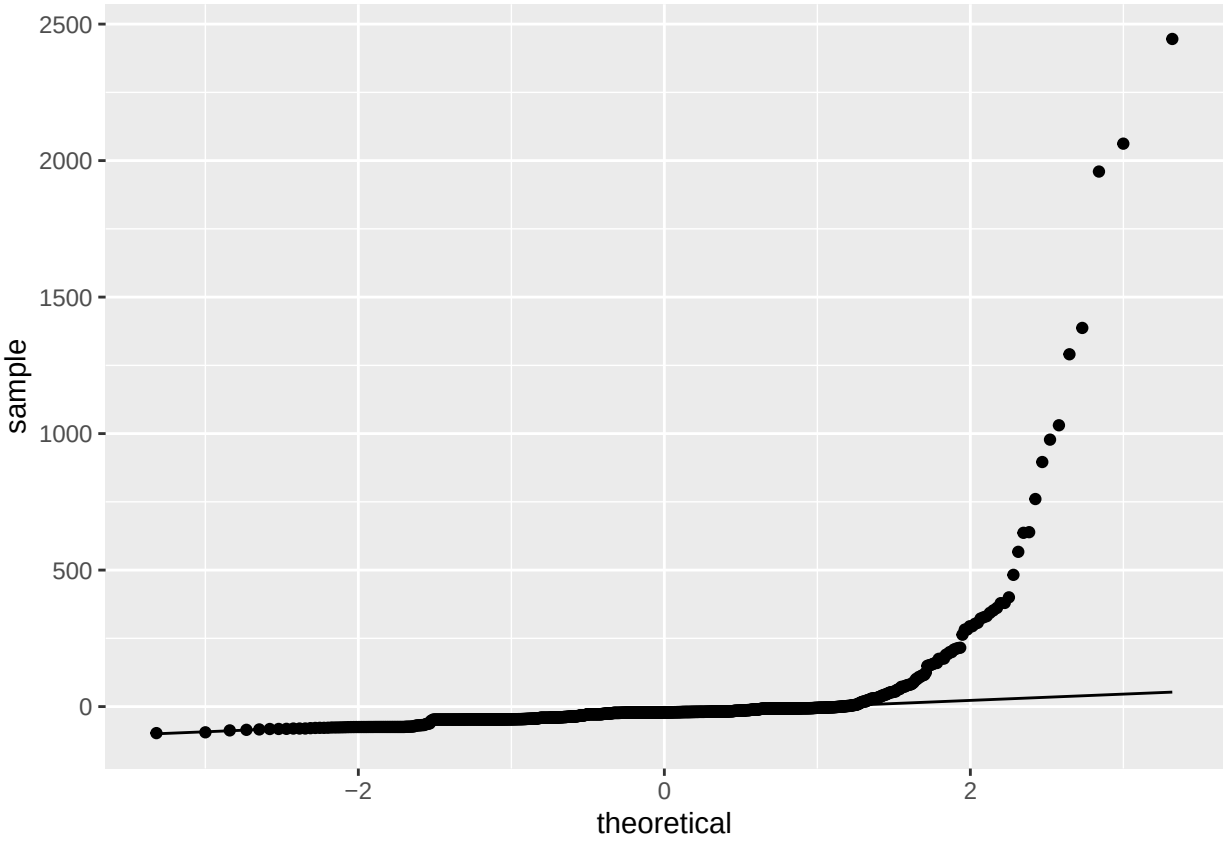
```
coefIntercept=rep(6.88246, times=15)
coefSlope=rep(-0.007810969, times=15)
for(i in 2:15)
{
  coefIntercept[i]=coefIntercept[i]+coefficients(LMwFE)[i]
  coefSlope[i]=coefSlope[i]+coefficients(LMwFE)[i+15]
}

plot(tnks~cd22, data=dat, main="Model v Orignal Data")
for(i in 1:15)
{
  abline(a=coefIntercept[i], b=coefSlope[i], col=(20+i), lty=2, lwd=2)
}
```




```
fit=fitted.values(LMwFE)
res=residuals(LMwFE)

p=ggplot(dat, aes(x=fit, y=res, color=subject.no))+
  geom_point()+
  geom_hline(yintercept = 0)+
  ggtitle("Fitted Vs Residual")+
  xlab("Fitted Value")+
  ylab("Residual Value")
p
```




```
q=ggplot(dat, aes(sample = res))+
  stat_qq()+
  stat_qq_line()
q
```



1. This model will be used primarily a a comparison, and it is not really the main object of interest, but I would

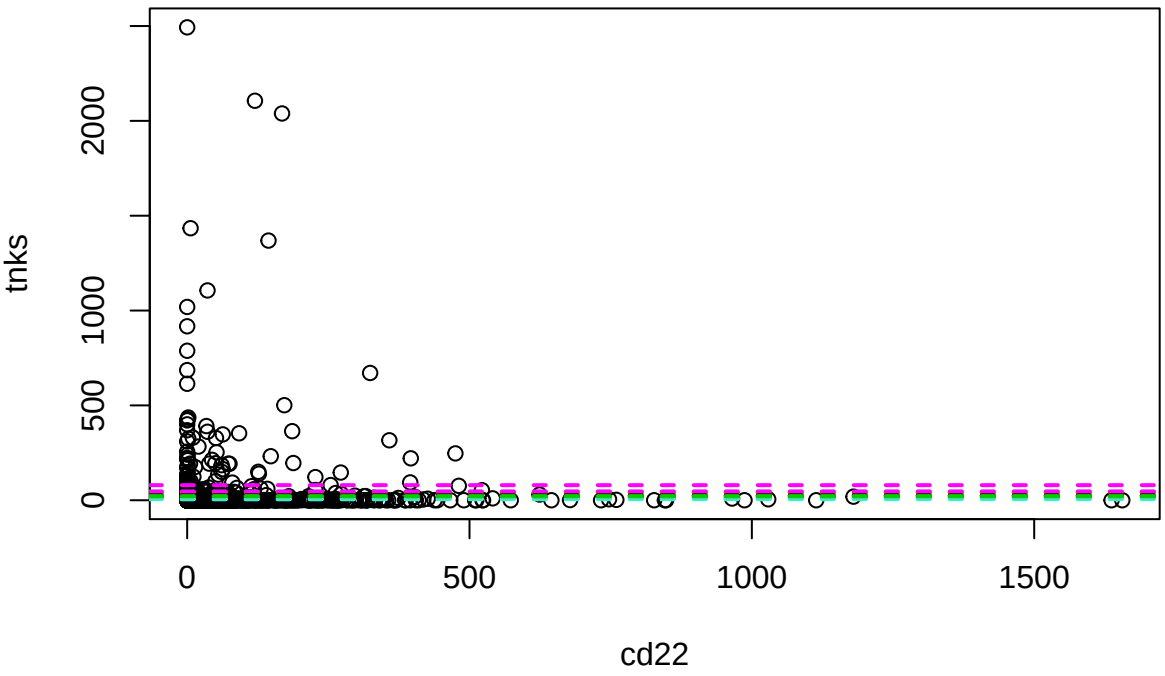
still like to know if I can incoperate the CD22 covariate into the model BY ITSELF, not in an interaction term with the subject.

2. I think that it will be important to be able to look back at this model and draw a comparative statistic (AIC, BIC, LR) for model comparison.

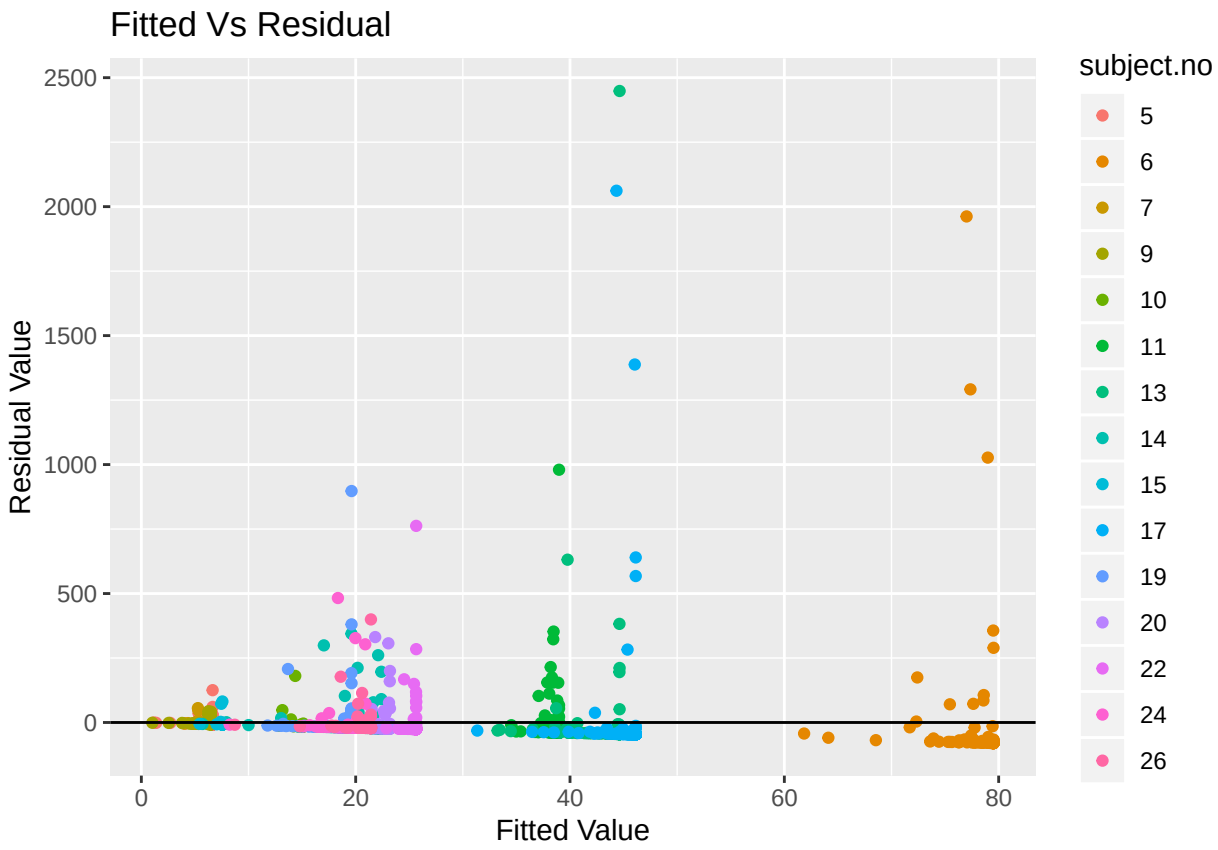
3. Next Step is further modeling

Model 2: Linear Mixed Model with Random Effect (LMMwRE)

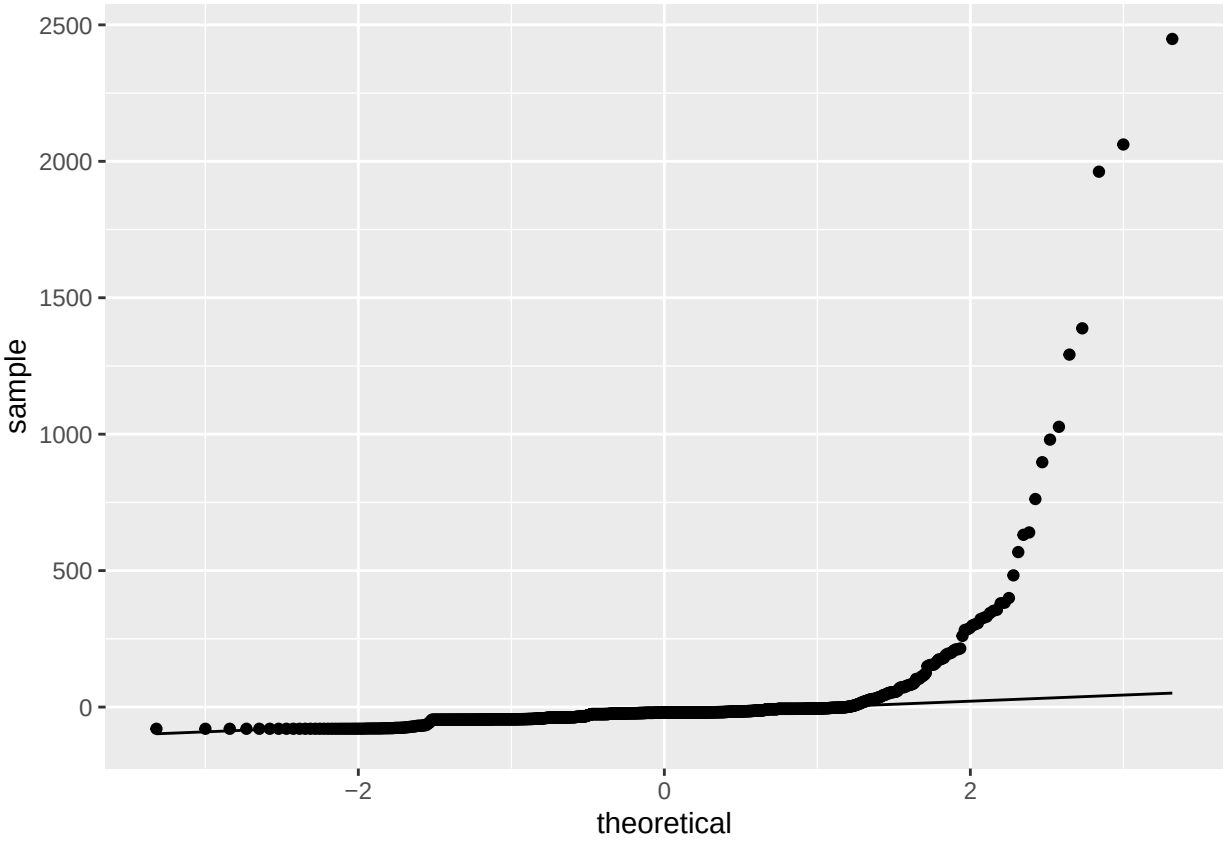
```
LMMwRE=lmer(tnks~subject.no+cd22+(1+cd22|subject.no), dat)
FixedEffects=fixef(LMMwRE)
RandomEffects=ranef(LMMwRE)

Intercept=rep(FixedEffects[1], times=15)
Slope=c()

for(i in 1:25){
  Slope[i]=RandomEffects[["subject.no"]][["cd22"]][i]
  Intercept[i]=Intercept[i]+RandomEffects[["subject.no"]][["(Intercept)"]][i]
}

for(i in 2:15)
{
  Intercept[i]=Intercept[i]+FixedEffects[i]
}

plot(tnks~cd22, data=dat, main="Model v Orignal Data")
for(i in 1:15)
{
  abline(a=Intercept[i], b=Slope[i], col=(20+i), lty=2, lwd=2)
}
```




```
fit=fitted.values(LMMwRE)
res=residuals(LMMwRE)

p=ggplot(dat, aes(x=fit, y=res, color=subject.no))+
  geom_point()+
  geom_hline(yintercept = 0)+
  ggtitle("Fitted Vs Residual")+
  xlab("Fitted Value")+
  ylab("Residual Value")
p
```




```
q=ggplot(dat, aes(sample = res))+
  stat_qq()+
  stat_qq_line()
q
```



1. I'm not completely sure I have captured the data structure accurately in these models. Specifically, I am

wondering if the theoretical models match the R models. I am also interested in whether I have accurately represented the clustering of the data.

2. Further considerations will be similar to that of the first model. I think that it should also be considered whether or not to do further outlier removal for the purpose of this project for visualization purposes.
3. Next step is to fit generalized linear models.

References

Arazi, Arnon, Deepak A Rao, Celine C Berthier, Anne Davidson, Yanyan Liu, Paul J Hoover, Adam Chicoine, et al. 2018. "The Immune Cell Landscape in Kidneys of Lupus Nephritis Patients." *bioRxiv*. Cold Spring Harbor Laboratory, 363051.

"Satija Lab." n.d. `https://satijalab.org/seurat/v3.0/pbmc3k_tutorial.html`.

"Zero-Inflated Model - Wikipedia." n.d. `https://en.wikipedia.org/wiki/Zero-inflated_model`.