Predicting drug residue depletion to establish a withdrawal period with data below the limit of quantitation (LOQ)

by

Yan McGowan

M.S., Kansas State University, 2012
M.B.A., Kansas State University, 2014

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Abstract

Veterinary drugs are used extensively for disease prevention and treatment in food producing animals. The residues of these drugs and their metabolites can pose risks for human health. Therefore, a withdrawal time is established to ensure consumer safety so that tissue, milk or eggs from treated animals cannot be harvested for human consumption until enough time has elapsed for the residue levels to decrease to safe concentrations. Part of the process to establish a withdrawal time involves a linear regression to model drug residue depletion over time. This regression model is used to calculate a one-sided, upper tolerance limit for the amount of drug residue remaining in target tissue as a function of time. The withdrawal period is then determined by finding the smallest time so that the upper tolerance limit falls below the maximum residue limit.

Observations with measured residue levels at or below the limit of quantitation (LOQ) of the analytical method present a special challenge in the estimation of the tolerance limit. Because values observed below the LOQ are thought to be unreliable, they add in an additional source of uncertainty and, if dealt with improperly or ignored, can introduce bias in the estimation of the withdrawal time. The U.S. Food and Drug Administration (FDA) suggests excluding such data while the European Medicine Agency (EMA) recommends replacing observations below the LOQ with a fixed number, specifically half the value of the LOQ. However, observations below LOQ are technically left censored and these methods are do not effectively address this fact. As an alternative, a regression method accounting for left-censoring is proposed and implemented in order to adequately model residue depletion over time. Furthermore, a method based on generalized (or fiducial) inference is developed to compute a tolerance limit with results from the proposed regression method. A simulation study is then conducted to compare the proposed

withdrawal time calculation procedure to the current FDA and EMA approaches. Finally, the proposed procedures are applied to real experimental data.

Predicting drug residue depletion to establish a withdrawal period with data below the limit of quantitation (LOQ)

by

Yan McGowan

M.S., Kansas State University, 2012
M.B.A., Kansas State University, 2014

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2019

Approved by:

Major Professor
Dr. Christopher I. Vahl

Copyright

Abstract

Veterinary drugs are used extensively for disease prevention and treatment in food producing animals. The residues of these drugs and their metabolites can pose risks for human health. Therefore, a withdrawal time is established to ensure consumer safety so that tissue, milk or eggs from treated animals cannot be harvested for human consumption until enough time has elapsed for the residue levels to decrease to safe concentrations. Part of the process to establish a withdrawal time involves a linear regression to model drug residue depletion over time. This regression model is used to calculate a one-sided, upper tolerance limit for the amount of drug residue remaining in target tissue as a function of time. The withdrawal period is then determined by finding the smallest time so that the upper tolerance limit falls below the maximum residue limit.

Observations with measured residue levels at or below the limit of quantitation (LOQ) of the analytical method present a special challenge in the estimation of the tolerance limit. Because values observed below the LOQ are thought to be unreliable, they add in an additional source of uncertainty and, if dealt with improperly or ignored, can introduce bias in the estimation of the withdrawal time. The U.S. Food and Drug Administration (FDA) suggests excluding such data while the European Medicine Agency (EMA) recommends replacing observations below the LOQ with a fixed number, specifically half the value of the LOQ. However, observations below LOQ are technically left censored and these methods are do not effectively address this fact. As an alternative, a regression method accounting for left-censoring is proposed and implemented in order to adequately model residue depletion over time. Furthermore, a method based on generalized (or fiducial) inference is developed to compute a tolerance limit with results from the proposed regression method. A simulation study is then conducted to compare the proposed

withdrawal time calculation procedure to the current FDA and EMA approaches. Finally, the proposed procedures are applied to real experimental data.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my appreciation to Dr. Christopher I. Vahl, my major professor, for all of his knowledge, guidance and suggestions. I learned a lot from him through many helpful discussions. He was always available when I have questions. My appreciation also goes to Dr. David G. Renter for his willingness to serve as the outside chairperson in the examining committee for my doctoral degree. I would also like to thank Dr. Wei-Wen Hsu, Dr. Gyuhyeong Goh and Dr. Steve Dritz for serving on my committee, and for their valuable guidance through my study process.

I would like to thank the department of statistics for offering me graduate assistantships so that I could come to the States and complete my graduate studies at Kansas State University. I would like to thank everyone in the department for their kindness, thank all the professors in the department for their excellent courses and for their help. Finally, many thanks to my family for their endless love, support, understanding and encouragement.

Chapter 1 - Introduction

The history of animal consumption by humans' dates back to 2.6 million years ago (Pobiner 2013). From then on, human beings started to demand more and more animals and animal products. In 2016 the total beef consumption in major food production countries (U.S.A., China, India, Brazil, etc.) reached 5.87 million metric tons (United States Department of Agriculture (USDA) Foreign Agricultural Service, 2017). Additionally, the total amount will increase by 1 million metric tons in 2017. In 2015, total U.S beef consumption was 24.8 billion pounds (USDA Economic Research Service, 2017a). However, beef is not the most highly consumed animal product. In the U.S. consumption of poultry and its products (mainly eggs) is considerably higher than for beef (USDA Economic Research Service, 2017b). The U.S. poultry industry is the world's largest producer and second largest exporter of poultry meat and is a major egg producer. More and more animal farms are founded to supply rapidly increasing demand for animal and animal products (egg, milk, etc.). At the same time, how to ensure supply by protecting animals from sickness and weakness has become a big question.

Veterinary medicine is a broad discipline that deals with disease prevention, diagnosis and treatment of animals. In order to help protect the health and welfare of animals and ensure food safety of animal products, such as egg, milk, meat, organs, etc., the food supply industry utilizes a range of veterinary drugs. These drugs are an important contributor to continuous industry growth. Administration of these drugs may result in drug residues in food. Residue is defined as "pharmacologically active substances (whether active principles, recipients or degradation products) and their metabolites which remain in foodstuffs obtained from animals to which the VMPs in question have been administered" by the European Medicines Agency

(EMA), a European Union agency responsible for the protection of public and animal health through the scientific evaluation and supervision of medicines (Beyene, 2016). Public health is directly related to the quality of food (Ames, 1983), and in particular to animals and animal products. The risks associated with veterinary drug residues that remain in the edible tissues of treated animals or animal products (egg and milk) can be a big health hazard to consumers.

Food safety of animals and animal products is a large concern of public health agencies around the world. Agencies such as EMA's Committee for Medicinal Products for Veterinary Use (CVMP) and the U.S. Food and Drug Administration's (FDA) Center for Veterinary Medicine (CVM regulate veterinary drugs to ensure the safety of animal products' consumption. When a veterinary drug gets approved, the drug label needs to show the active ingredients, identification of the animal(s) to be treated, adequate directions for proper use, and cautions/precautions including milk and meat withdrawal times, etc. Normally, the application of veterinary drugs has to follow the labels' guidelines. However, there is an exception, extra-label drug use (ELDU). An ELDU refers to the use of an approved drug in a manner that is not in accordance with the approved label directions, including drug use in other species, use with different a dose or frequency, or use via a different route of administration. But all ELDU must follow the conditions set forth by the Animal Medicinal Drug Use Clarification Act of 1994 (AMDUCA) and the U.S. Food and Drug Administration (FDA) regulations. Veterinarians must assure that the identity of the treated animal(s) is carefully maintained and establish a substantially extended withdrawal period supported by appropriate scientific information prior to marketing milk, meat, eggs, or other edible products from the treated animals(s). Predicting the concentration of drug residue is an important procedure to establish withdrawal period.

Withdrawal times are established to ensure consumer safety, which means that edible tissue, milk or eggs cannot be harvested for human consumption until enough time has elapsed after treatment for residue levels to deplete to safe concentrations. Since withdrawal periods are very important in veterinary medicine research to ensure public health, the calculation procedure has been well established by the FDA, EMA and other agencies, but with a slightly different tolerance interval setting. However, it is common to have drug residue data below the limit of quantitation (LOQ) and/or the limit of detection (LOD) due to the measurement limitations of analytical instruments. The LOQ is the lowest concentration that can be reliably detected and at which some predefined goals for bias and imprecision are met. On the other hand, LOD is the lowest concentration that can be distinguished from the highest apparent analyte concentration expected to be found when replicates of a blank sample containing no analyte are tested. Finding an adequate approach to analyze data at and below the LOQ and/or LOD still obsesses regulatory officers and professional researchers.

We propose a left censored data regression to estimate the withdrawal period with data that fall below the LOQ. We will also compare the withdrawal period with this approach to those estimated using the recommended methods from the FDA and EMA. In addition, we will apply this approach to the real experiment data which can be found in EMA-CVMP (2016).

1.1 Limit of quantitation (LOQ)

In drug and chemical studies, the purpose of an analytical method is the delivery of a qualitative and/or quantitative result with an acceptable uncertainty level (Şengül 2016). The minimum concentrations of an analyte that can be realizably detected or measured in an analytical procedure are important performance indicators of certain assays (Lawson 1994). In practice, precision, trueness, selectivity/specificity, linearity, operating range, recovery, limit of blank (LOB), limit of detection (LOD), limit of quantification (LOQ), sensitivity, ruggedness/robustness, and applicability are all terms used to describe the smallest concentration of an analyte that can be realizably measured by the analytical procedure (Armbruster and Pry 2008). These terms are important not only to evaluate the analytical performance of each laboratory but also to determine dynamic range or analytical measurement range. In addition, clinical laboratorians may deal with the analytical issue of when the ability of a laboratory test to detect a small amount of analyte is clinically significant. For example, this situation can arise when the medical decision level (such as maximum residual level) is at or below the analytical limits. It means the clinical action will depend on measurements of low concentration. In ELDU studies in certain tissues such as eggs, the FDA requires a zero-drug concentration withdrawal period. This analysis relies on low concentration measurements. LOB, LOD and LOQ are widely used by analytical chemists that test for abused drugs (Lawson 1994).

In 2004, the Clinical and Laboratory Institution (CLSI) published “Protocol for Determination of Limits of Detection and Limit of Quantitation” to standardize methods for determining LOB, LOD and LOQ (CLSI 2004). The LOB is the highest apparent analyte concentration expected to be found in replicates of a sample containing no analyte. It can be

described by $LOB = mean_{blank} + 1.645SD_{blank}$, where SD is standard deviation and 1.645 is the upper 0.05 critical value from a standard normal distribution. LOB is the upper 95% confidence limit of blank samples. For analytical procedures, the LOD is the threshold below which measurements are not significantly different from a blank signal with a given probability (Armbruster, Tillman and Hubbs 1994). It is the lowest analyte concentration to be reliably distinguished from LOB and at which detection is feasible. As defined in EP17, the LOD can be described by $LOB = LOB + 1.645SD_{low\ concentration\ sample}$ (CLSI 2004), where SD is the standard deviation of replicate tests of samples known to contain low concentrations of the analyte. LOQ is the lowest concentration of analyte can be measured with certain pre-defined precision at which the analyte cannot only be reliably detected but also at which some goals for bias and imprecision are met. It can be expected that the LOQ is equal to or larger than LOD but cannot be lower than LOD. Even though data below LOQ cannot meet certain precision criteria, there is some information within the data defined to be below LOQ. Therefore, we should find a way to deal with such data, especially for research based on low concentration data.

Based on FDA-CVM (2016), in the US FDA suggests eliminating data below LOD and LOQ. The EMA recommends using half of the LOQ to replace measurements that fall below the LOQ (EMA-CVMP 2016). In fact, there are more methods to handle situations where data fall below the LOQ. Using data that fall below the LOQ as measured, without any adjustment, has been demonstrated to be a bad approach (EMA-CVMP 2016). Maximum likelihood estimates, which determine the depletion curve that would maximize the likelihood of the observed data is also not recommended, based on the results of a simulation study by the EMA (EMA-CVMP 2016). Approaches using order statistics have been studied by Yan (2014). There are additional approaches to dealing with data below LOQ (Beal 2001; Senn and Hockey 2012): discard LOQ

data and estimate using remaining values as if they came from a full distribution; discard LOQ data and estimate by treating the remaining values as forming a 'truncated' sample; calculate the likelihood of all remaining samples conditional on the value being greater than the LOQ; ignore any actual values of the LOQ data and estimate by treating the sample as a whole as one in which LOQ values are censored; calculate the likelihood of the LOQ sample assuming that the value is less than the LOQ; estimate the likelihood as in last method but add an additional constraint that all LOQ values must be positive; calculate the likelihood of all values conditional on their being greater than zero with the additional constraint for LOQ values that they are less than the LOQ; impute LOQ data by one-half the LOQ and estimate as if all the values were real; when measurements are taken for a given individual over time, impute as for last method for the first LOQ measurement and discard all subsequent LOQ data; impute LOQ values by zero and estimate as if all the values were real.

Censored data has been studied extensively in environmental research (Helsel 2011). This dissertation proposes using left censored data regression to handle data below LOQ. The details are discussed in chapter 2.

1.2 Withdrawal Period

The withdrawal period is the necessary interval between the last administration of the drug under normal conditions of use and the time when treated animals or animal products can be consumed by the public. The U.S. Food and Drug Agency's (FDA) Center for Veterinary Medicine (CVM) regulates the evaluation of every veterinary drug in each food animal species for human food safety. The withdrawal period is one of the important steps in this evaluation approach. A detailed flow diagram of food safety evaluation approach is shown in Figure 1.1 (FDA 1994; FDA-CVM 2016). All legally approved drugs are evaluated by this process. Withdrawal period can be established for specific marker residues in a given target tissue.

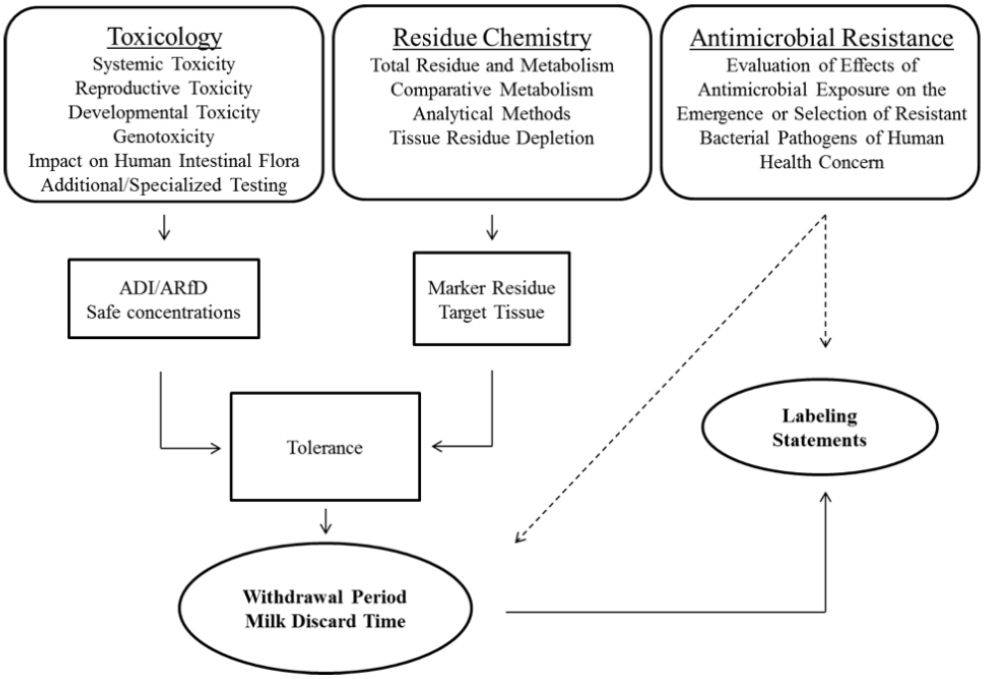

Figure 1.1: Flow Diagram Illustrating the General Approach for the Human Food Safety Evaluation. The dashed lines indicate that the component applies to new animal drugs with antimicrobial activity. ADI means Acceptable Daily Intake. ARfD means Acute Reference Dose. (FDA-CVM, 2016)

However, for extra-label drug use, there is no legally proved withdrawal period to follow for consumers, researchers, backyard farmers, etc. Hence, estimating withdrawal period becomes very important for extra-label veterinary drug use to make sure edible animal product are safe since there is no official guided information. This shows the importance of our research since some drugs are used extra-label and no withdrawal period has been established for its marker residue.

For regular data, which does not contain data below LOD and LOQ, the EMA's Committee for Medicinal Products for Veterinary Use (CVMP) and the U.S. Food and Drug Agency (FDA)'s Center for Veterinary Medicine (CVM) provide clear regulatory guidelines to establish drug withdrawal periods. Here is an example to show how to establish withdrawal period using regular data. Following FDA computation recommendations, a veterinary drug withdrawal period in liver tissue was established with an R program provided by CVM. The withdrawal period is the time when the upper one-sided tolerance limit with a given confidence is below a certain limit. Normally, we use the maximum residue limit (MRL) as this certain critical limit. Assume a limit of 5.2 ppm was used in this case based on FDA guidelines for a drug marker residue in target tissue. From the Figure 1.2, it is clear that the withdraw period is 24 hours. It was rounded up from 23.6689 hour, since the FDA and EMA require to rounding up to next day when the calculated withdrawal period is not a full day.

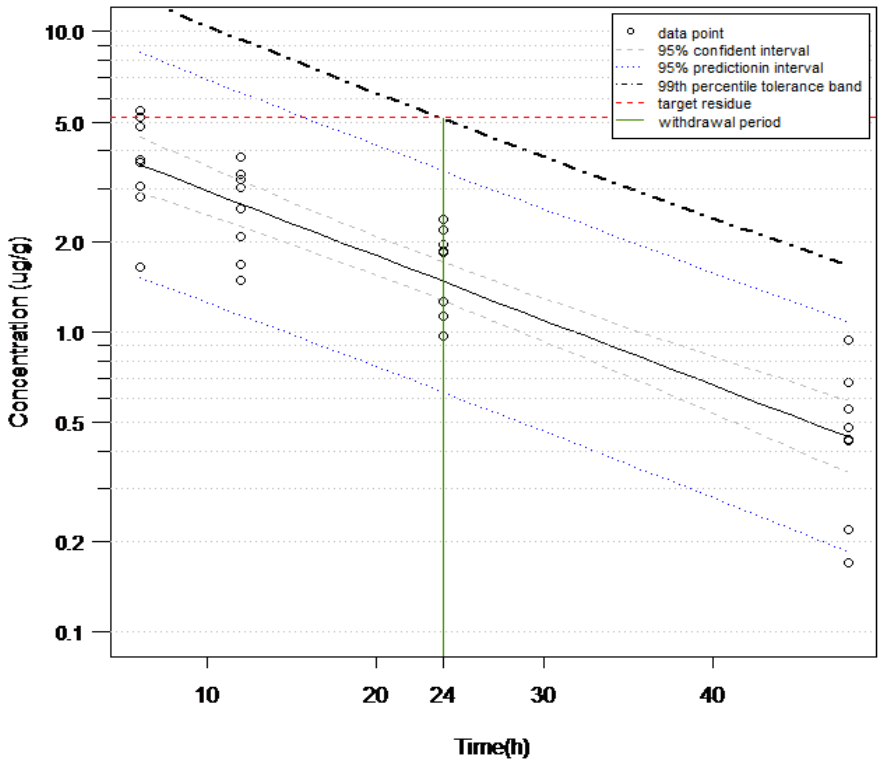

Figure 1.2: The drug residue depletion regression and withdrawal period estimate in example. The withdrawal period is 24 hours.

From this example, it is clear that, in order to construct withdrawal periods, we have to establish linear regression using experimental data and calculate a regression tolerance band. Establishing the MRL is also very important; however, this dissertation focuses on statistical model and method development for the estimation of withdrawal periods that can be applied using any value of MLR. Establishing the MRL requires professional drug and toxicology knowledge and thus tends to fall within veterinary research rather than statistical. In order to focus on the statistical aspects of estimating a withdrawal period, this study uses 5.2 ppm as the MRL. Here the 99^{th} percentile tolerance band is 99% upper tolerance limit with 95% confidence level (FDA 1994; FDA-CVM 2016). This concept will be explained in more detail in Section

1.5. Moreover, we will develop new approach to establish the tolerance limit; this will be shown in Chapter 3.

Since our data contains observations below LOQ, a left-censored data linear regression was estimated to handle LOQ issue. The new tolerance limit approach will be applied to this regression to establish 99% upper tolerance limit with 95% confidence level. This is slightly different from the tolerance limit calculated using EMA and FDA guidance (EMA-CVMP 2016; FDA-CVM 2016) which will be explained in Section 1.5.

1.3 Linear Regression to Predict Drug Residue Over Time

It is clear that withdrawal time is the time at the point of intersection between the target residue concentration and the upper 99th percentile tolerance band (99% tolerance limit with a 95% confidence level based on FDA guidance). In order to establish the withdrawal period, besides determining the critical drug residue limit, the other important step is estimation of the upper tolerance limit. According to the pharmacokinetic compartment model theory, the relationship between drug concentration and time through all phases of absorption, distribution and elimination is usually described by multi-exponential mathematical terms (EMA-CVMP 2016),

$$C_t = \sum_{i=1}^{n} C_{0,i} e^{-\lambda_i t}$$

where C_t is the concentration at time t, $C_{0,i}$ is the extrapolated concentration at time $t = 0$, and i^{th} exponential term, and λ_i is the constant rate corresponding to the i^{th} exponential term. Here the exponential term is the compartment. In the pharmacokinetic compartment model, the compartments are tissues or function organs. Here, the compartment can be viewed as a target tissue.

The FDA's CVM makes a simplifying assumption for the analysis of residue data that the depletion curve, during the phase of the depletion closest to the established tolerance, can be represented by a single exponential equation. The EMA-CVMP (2016) also claims that with a one compartment model,

$$C_t = C_0 e^{-\lambda t} \ ,$$

linearity of the plot $\ln(C)$ versus time indicates that the model for residue depletion is applicable and linear regression analysis of the logarithmic transformed data can be considered for the calculation of withdrawal periods. So, the natural logarithm transformation of concentration can be fitted in a linear regression, which

$$\ln(C_t) = \alpha - \lambda t$$

where $\ln(C_t)$ is the log of concentration at time *t* and α is the intercept in which $\alpha = \ln(C_0)$. The term λ is the rate constant. Hence, EMA and FDA use linear regression to predict drug residue depletion over time with a logarithmic transformation of concentration.

Based on a lognormal assumption, it is clear that the logarithmic transformation of concentration follows normal distribution. Hence, the left censored data regression model developed in this study focuses on cases with the assumption of a normal distribution.

1.4 Tolerance Interval

Regression models have been used for prediction for a wide range of purposes. However, only reporting the predictions is not satisfactory. Statistical intervals are needed to quantify the uncertainty about the scalar quantity. There are many types of intervals available, such as confidence intervals, prediction intervals, tolerance intervals, and so on. The method to choose a certain interval depends on the problem, assumptions and application (De Gryze, Langhans, & M. Vandebroek, 2007). Several measures of statistical intervals (limits) are used to describe the possible range of values given the variability among the observations (FDA-CVM 2016). However, there are certain disadvantages with these intervals (limits) (Myers 1990). For example, the commonly used 95% confidence interval means we are 95% confidence data falls into the interval at time t, which does not mean 95% of the data falls into the interval. The confidence interval provides a range of values that we can believe, with a given level of confidence, contains the true value of a variable in the larger population that is being estimated by the sample in a particular study (Hahn & Meeker, 2001). However, the probability coverage is not 95%, which means the 95% confidence interval does not indicate that there is a 95% probability that the population parameter lies within the interval, i.e. a 95% probability that the interval covers the population parameter (Rasmussen, Staggs, Beard & Newman, 1998; De Gryze et al., 2007). On the other hand, prediction intervals consider the accuracy with which we can predict the targets themselves (Heskes, 1997). The probability coverage of the 95% prediction interval is also not 95% (De Gryze et al., 2007).

Tolerance intervals can provide a range that contains a specific proportion or more of the sampled population (Krishnamoorthy & Mathew, 2009), which means the probability coverage

of a 95% tolerance interval is 95% with certain level of confidence. Sample size is relevantly small in drug residue research due to financial, labor and equipment limitations. The tolerance interval also works great well with small sample sizes (De Gryze et al., 2007).

De Gryze et al. (2007) denoted the tolerance interval at x_{n+1}, which contains a proportion p of the individual responses $y|x_{n+1}$ with confidence $1-\alpha$ is defined as the interval $I_T(x_{n+1})$ around the predicted response $\hat{y}|x_{n+1}$ for which

$$P_r\left(P_r\left(y\middle|x_{n+1} \in I_T(x_{n+1})\right) \geq p\right) = 1-\alpha$$

There are two probability parts in in the tolerance interval. The probability $P_r\left(y\middle|x_{n+1} \in I_T(x_{n+1})\right)$ is called the coverage of the interval $I_T(x_{n+1})$. Hence, the tolerance interval can cover a fixed proportion of the population. Based on the definition equation, the upper tolerance limit can be computed subject to the condition that at least 95% of the population response levels are below the limit, with a certain confidence level $1-\alpha$.

The FDA's CVM suggests 99% setting for the proportion p with 95% confidence level. In another words, the FDA requires the tolerance limits to cover 99% population with 95% confidence. There is a slight difference in the EMA guidance. Instead of 99% setting for proportion p, a 95% population coverage with 95% confidence is required to calculate the tolerance limit. The tolerance limit is the one side boundary of the tolerance interval. In drug withdrawal period studies, we focus on the upper boundary of tolerance interval. The general tolerance limit, FDA tolerance limit and EMA tolerance limit will be explained in detail in chapter 3.

Chapter 2 - Left Censored Data Regression

In chemical concentration data studies, it is very common to have values that are below defined thresholds such as the LOQ and LOD (USEPA 2009; Boeckel et al., 2015; Ichihara et al., 2017). Such data sets are viewed as censored data. Censored data sets have unknown values beyond a limit on either end of the number line. In other words, a series of samples contains censored data if some of the observations that are below or above a specific threshold value. However, censored data is different from truncated data. In truncated data, all values that exceed some limit are not recorded. In this dissertation, we focus on data below LOQ, which will be considered to be censored data and not truncated.

Even though measurements below the LOQ are less accurate, they still contain information. At a minimum, we know these data fall below a certain level. Statistical methods for dealing with censored data have a long and strong history in survival analysis and life testing research (Kalbfleisch & Prentice 1980; Glasziou, Simes & Gelber 1990; Kelly& Lim 2000; Miller 2011; Cox 2018). Survival analysis is focused mainly on right censored data analysis since the censored observations are above certain time limit. Our research is different from survival data in that it is focused on observations that contain data below certain level, such as LOQ.

In order to predict drug residue depletion process during time, a left-censored data regression model was developed to handle data below LOQ issue in this dissertation. The details are introduced in the following sections.

2.1 Likelihood Function

In the most general case for censored data, the likelihood function is the product of two components: one for the censored data and another for uncensored data. Based on this theory, Aitkin (1981) used function (2.1) to define the likelihood for right censored survival data by considering *n* items as detected data and *m* items that survived until a cut off time point. That means there are m observations which are censored because of the limitation of clinical trials. Observations y_i , $(i = 1, 2, 3, \cdots, n, n+1, n+2, \cdots n+m)$, are independent and normal distributed lifetimes with mean μ_i and variance σ^2.

$$L = \frac{1}{\sigma^n} \prod_{i=1}^{n} \phi(z_i) \prod_{i=n+1}^{m} \psi(z_i) \qquad (2.1)$$

Where, $z_i = \frac{(y_i - \mu_i)}{\sigma}$,

$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-1/2y^2}$ and

$\psi(y) = \int_y^{\infty} \phi(t)dt.$

In 2011, Hesel used this concept to develop likelihood for censored data. He summarized the likelihood function for censored data as the product of three parts. One part is still for detected observations. Here, the censored part includes two parts: one represents for left censored data and the other for right censored data. Hence, the likelihood function can be written as

$$L = \prod \prod p(x) \prod \prod F(x) \prod \prod S(x) \qquad (2.2)$$

where, $p(x)$ is the probability density function of the uncensored data. It indicates the frequency of having detected observation x. $F(x)$ and $S(x)$ are both for the censored data. $S(x)$ is survival function, which shows the probability of observations exceeding value x. Observations greater than value x are called right censored data; thus, $S(x)$ is determined by right censored observations. $F(x)$ is very similar to $S(x)$. Instead of showing the probability of observations exceeding value x, $F(x)$ describes the probability of observations equal to or less than value x. This portion of data is called left censored observations. $F(x)$ is determined by left censored observations. The relationship between $F(x)$ and $S(x)$ is $F(x) = 1 - S(x)$.

In our study, below limit of quantitation (BLOQ) data is left censored data since it is data below limit of quantitation. Based on previous studies, this research defines the likelihood function for data including left censored observations as

$$L = \prod_{i=1}^{n} f(y_i)^{\tau_i} F(y_i)^{1-\tau_i} \qquad (2.3)$$

where, $\mathrm{i} = 1, 2, 3, \cdots, n$

$f(y_i)$ is the probability density function of detected data, which are the observations above the LOQ,

$F(y_i)$ is the cumulative density function of censored data, which are the observations equal to or less than the LOQ,

τ_i is an indicator variable where τ_i equals 1 when observation i is above the LOQ and τ_i equals 0 when observation i is equal to or less than the LOQ

Now, we assume observations y_i are independent and normally distributed with mean μ_i and common variance σ^2, then the likelihood function can be defined as

$$L = \prod_{i=1}^{n} f(y_i)^{\tau_i} \Phi\left(\frac{a-\mu_i}{\sigma}\right)^{1-\tau_i} \qquad (2.4)$$

where, $a = LOQ$,

$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\mu_i)^2}{2\sigma^2}}$,

$F(y_i) = \int_0^{LOQ} f(t)dt = \Phi\left(\frac{LOQ-\mu_i}{\sigma}\right) = \Phi\left(\frac{a-\mu_i}{\sigma}\right)$, which is a cumulative density function (CDF) of a standard normal distribution $z_i = \left(\frac{a-\mu_i}{\sigma}\right)$

τ_i = 0, when observation y_i is equal to or less than the LOQ, which is left censored

τ_i = 1, when observation y_i is above the LOQ, which is detected

Compared to the methods from Aitkin (1981) and Thompson and Nelson (2003), the big advantage of this proposed likelihood function is reduced data pre-process steps: it is not necessary to sort the data set first, all of the censored and uncensored data can be pooled together, and the analysis can be done with the whole dataset after setting a threshold number, which in our study is the LOQ.

2.2 Maximum likelihood estimations

Maximum likelihood method is one method to calculate the desired estimates (Haley & Knott 1992). In addition, maximum likelihood estimation (MLE) is increasingly used in a variety of research (Bańbura & Modugno, 2014; Efron 2018; Améndola, Drton & Sturmfels, 2015). When the distribution can be assumed from prior knowledge outside the dataset, the MLE method can perform very well even with small sample sizes (Helsel 2011). In this dissertation, we know the natural logarithmic transformation of drug residue concentration should follow normal distribution. It is therefore appropriate to use MLE in this research.

Suppose Y is a dependent variable with observations y_i (i $= 1, 2, 3, \cdots, n$), where y_i can be an observed detected value or an inaccurate number below LOQ. In addition, the y_i terms are independent from each other and normally distributed with mean μ_i and common variance σ^2. The X_j terms are explanatory variables with observations x_{ij} for the j^{th} explanatory variable and i^{th} observation item. Then in linear regression setting, with $x_{i0} = 1$, we can have

$$\mu_i = \sum_{j=0}^{k} \beta_j x_{ij} \qquad (2.5)$$

Now, based on function (2.4), the likelihood function L in a linear regression model for left censored data can be described by

$$L = \prod_{i=1}^{n} f(y_i)^{\tau_i} \Phi\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)^{1-\tau_i} \qquad (2.6)$$

where, $a = LOQ$,

$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \sum_{j=0}^{k} \beta_j x_{ij})^2}{2\sigma^2}}$,

$\Phi\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$ is a CDF of a standard normal distribution which $z_i = \left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$

τ_i =0, when response variable observation y_i is equal to or less than the LOQ,

τ_i =1, when response variable observation y_i is above the LOQ

MLE is commonly used in the survival analysis (right censored data) to estimate parameters (Wingo 1993; Bhattacharyya 1985; Miller 2011). It is based on the likelihood function L. By solving $\arg\max_{\theta \in \Theta} L(\theta)$, we can get MLE of parameter θ, the θ which maximizes likelihood function. In practice, it is very convenient to work with the natural logarithm of the likelihood (log-likelihood) function instead likelihood function. Then the MLE of θ can be easily calculated by solving equation $\frac{\partial \log L(\theta)}{\partial \theta} = 0$.

Here, the log-likelihood function for left censored data in a linear regression setting should be

$$\log L = \sum_{i=1}^{n} \tau_i \log f(y_i) + (1 - \tau_i) \log \Phi(z_i) \qquad (2.7)$$

where, $f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i - \sum_{j=0}^{k} \beta_j x_{ij})^2}{2\sigma^2}}$, $z_i = \left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$

$\log f(y_i) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2$ and

$\log \Phi(z_i) = \log \Phi\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$.

Then maximizing the log-likelihood function can be achieved by taking the partial derivative of $\log L$ with respect to parameters β_j and σ equals to zero. That means in order to get MLE of β_j, we need to have $\frac{\partial \log L}{\partial \beta_j} = 0$. On the other hand, by solving equation $\frac{\partial \log L}{\partial \sigma} = 0$, MLE of σ can be obtained. Steps in detail are showed below.

For parameter β_j, in order to get MLE of β_j, we can solve the function (2.8)

$$0 = \frac{\partial \log L}{\partial \beta_j} = \frac{\sum_{i=1}^{n} \tau_i \log f(y_i) + (1-\tau_i) \log \Phi(z_i)}{\partial \beta_j} \qquad (2.8)$$

In order to make calculation easier, let's compute $\frac{\partial \log f(y_i)}{\partial \beta_j}$ and $\frac{\partial \log \Phi(z_i)}{\partial \beta_j}$ first.

$$\frac{\partial \log f(y_i)}{\partial \beta_j} = \frac{\partial \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2}{2\sigma^2}}}{\partial \beta_j}$$

$$= \frac{\partial -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2}{\partial \beta_j}$$

$$= \frac{1}{\sigma^2}\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right) \cdot x_{ij} \qquad (2.9)$$

$$\frac{\partial \log \Phi(z_i)}{\partial \beta_j} = \frac{\partial \log \Phi\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)}{\partial \beta_j}$$

$$= \frac{\partial \Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) \Big/ \partial \beta_j}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}$$

$$= \frac{-\frac{1}{\sigma} f\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) \cdot x_{ij}}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}$$

$$= \frac{1}{\sigma^2} \cdot \left(-\sigma \cdot \frac{f\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)} \cdot x_{ij} \right)$$

$$= \frac{1}{\sigma^2} \left(-\sigma S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) + \sum_{j=0}^{k}\beta_j x_{ij} - \sum_{j=0}^{k}\beta_j x_{ij} \right) \cdot x_{ij}$$

$$= \frac{1}{\sigma^2}(w_i - \mu_i) \cdot x_{ij} \qquad (2.10)$$

where, $S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) = \frac{f\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}$ and $w_i = \sum_{j=0}^{k}\beta_j x_{ij} - \sigma S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)$.

By citing function (2.9) and (2.10), function (2.8) can be extended to function (2.11)

$$0 = \frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{n} \tau_i \frac{1}{\sigma^2}\left(y_i - \sum_{j=0}^{k}\beta_j x_{ij} \right) \cdot x_{ij} + (1-\tau_i)\frac{1}{\sigma^2}\left(w_i - \sum_{j=0}^{k}\beta_j x_{ij} \right) \cdot x_{ij}$$

$$= \sum_{i=1}^{n}\left(y_i^* - \sum_{j=0}^{k}\beta_j x_{ij} \right) \cdot x_{ij} \qquad (2.11)$$

where, $y_i^* = \begin{cases} y_i, \ \tau_i = 1 \ when \ y_i \ is \ above \ LOQ \ value \ a \\ w_i, \ \tau_i = 0 \ when \ y_i \ is \ equal \ to \ or \ less \ than \ LOQ \ value \ a \end{cases}$

So, the MLE of parameter β_j is equivalent to the least square estimate of β_j with y_i^*, which the response variable y_i^* are censored observations $w_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ij} - \hat{\sigma} S\left(\frac{a - \sum_{j=0}^{k} \hat{\beta}_j x_{ij}}{\hat{\sigma}}\right)$ and original detected observations y_i.

Further, in order to get the MLE of σ, we can solve the function (2.12), shown below

$$0 = \frac{\partial \log L}{\partial \sigma} = \frac{\sum_{i=1}^{n} \tau_i \log f(y_i) + (1 - \tau_i) \log \Phi(z_i)}{\partial \sigma} \qquad (2.12)$$

Similar to the calculation process for the MLE of β, we compute $\frac{\partial \log f(y_i)}{\partial \sigma}$ and $\frac{\partial \log \Phi(z_i)}{\partial \sigma}$ first.

$$\frac{\partial \log f(y_i)}{\partial \sigma} = \frac{\partial \log \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2}{2\sigma^2}}}{\partial \sigma}$$

$$= \frac{\partial -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2}{\partial \sigma}$$

$$= -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(y_i - \sum_{j=0}^{k} \beta_j x_{ij}\right)^2 \qquad (2.13)$$

$$\frac{\partial \log \Phi(z_i)}{\partial \sigma} = \frac{\partial \log \Phi\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)}{\partial \sigma}$$

$$= \frac{\partial \Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)/\partial\sigma}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}$$

$$= \frac{f\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)} \cdot \frac{-\left(a-\sum_{j=0}^{k}\beta_j x_{ij}\right)}{\sigma^2}$$

$$= \frac{-\left(a-\sum_{j=0}^{k}\beta_j x_{ij}\right)}{\sigma} \cdot \frac{1}{\sigma} \cdot S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) \qquad (2.14)$$

where, $S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right) = \frac{f\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}{\Phi\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)}$.

Hence, by using function (2.13) and function (2.14), function (2.12) can be written as

$$0 = \frac{\partial \log L}{\partial\sigma} = \sum_{i=1}^{n} \tau_i \left(-\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(y_i - \sum_{j=0}^{k}\beta_j x_{ij}\right)^2 \right)$$

$$+(1-\tau_i)\frac{-\left(a-\sum_{j=0}^{k}\beta_j x_{ij}\right)}{\sigma} \cdot \frac{1}{\sigma} \cdot S\left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)$$

$$= \sum_{i=1}^{n} -\frac{\tau_i}{\sigma} + \frac{\tau_i}{\sigma^3}\left(y_i - \sum_{j=0}^{k}\beta_j x_{ij}\right)^2 - (1-\tau_i) z_i \cdot \frac{1}{\sigma} \cdot S(z_i)$$

$$= -\frac{\sum_{i=1}^{n}\tau_i + \sum_{i=1}^{n}(1-\tau_i) z_i S(z_i)}{\sigma^3}$$

$$\cdot \left(\frac{\sum_{i=1}^{n}\tau_i \left(y_i - \sum_{j=0}^{k}\beta_j x_{ij}\right)^2}{\sum_{i=1}^{n}\tau_i + \sum_{i=1}^{n}(1-\tau_i) z_i S(z_i)} - \sigma^2 \right) \qquad (2.15)$$

where, $z_i = \left(\frac{a-\sum_{j=0}^{k}\beta_j x_{ij}}{\sigma}\right)$.

Then, by solving function (2.15), we can calculate MLE of σ^2 with

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n}\tau_i\left(y_i - \sum_{j=0}^{k}\hat{\beta}_j x_{ij}\right)^2}{\sum_{i=1}^{n}\tau_i + \sum_{i=1}^{n}(1-\tau_i)\hat{z}_i S(\hat{z}_i)} \qquad (2.16)$$

where, $\hat{z}_i = \left(\frac{a-\sum_{j=0}^{k}\widehat{\beta}_j x_{ij}}{\widehat{\sigma}}\right)$.

2.3 EM algorithm

It is easy to see $\hat{\beta}_j$ and $\widehat{\sigma^2}$ can't be calculated directly based on function (2.11) and (2.16). In order to solve this problem, the expectation–maximization (EM) algorithm will be introduced in this section. The EM algorithm is an iterative computational method which is used to find MLE of parameters when equations can't be solved directly. Dempster, Laird and Rubin (1977) originally explained and established the EM algorithm for maximum likelihood estimations. The EM algorithm contains two steps: an expectation (E) step and a maximization (M) step. In general, the E step is to calculate conditional expectation of the complete log likelihood given current estimated parameters from M step. At the same time, the M step finds estimates, which maximize the expected complete log likelihood calculated from E-step. Then E step and M step will be repeated until the result can pass certain criterion.

In our study, we can get MLE using the following iterative method. First, we set the initial estimates of β_j and σ^2 by using all observations as uncensored data. We pretend censored observations (equal to or less than LOQ) are detected observations. Then the new response variable y_i^* can be established by the original detected observations y_i and $w_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ij_i} - \hat{\sigma} S\left(\frac{a-\sum_{j=0}^{k} \hat{\beta}_j x_{ij}}{\hat{\sigma}}\right)$. The new estimate of β_j will be least square estimate by solving function (2.11), which $\sum_{i=1}^{n}\left(y_i^* - \sum_{j=0}^{k} \beta_j x_{ij}\right) \cdot x_{ij} = 0$. Then we can compute $\hat{z}_i$ by $\hat{z}_i = \left(\frac{a-\sum_{j=0}^{k} \hat{\beta}_j x_{ij}}{\hat{\sigma}}\right)$.

Based on $\hat{z}_i$ and $\hat{\beta}_j$, $\widehat{\sigma^2}$ can be obtained by $\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \tau_i \left(y_i - \sum_{j=0}^{k} \hat{\beta}_j x_{ij}\right)^2}{\sum_{i=1}^{n} \tau_i + \sum_{i=1}^{n} (1-\tau_i) \hat{z}_i S(\hat{z}_i)}$. Then the new estimation of parameters $\hat{\beta}_j$, $\widehat{\sigma^2}$ will be inserted into calculation function $w_i = \sum_{j=0}^{k} \hat{\beta}_j x_{ij_i} -$

$\hat{\sigma}S\left(\frac{a-\sum_{j=0}^{k}\hat{\beta}_j x_{ij}}{\hat{\sigma}}\right)$ to generate response variable y_i^* . This iterative process will keep going until convergence.

Based on the EM algorithm theory from Dempster, Laird and Rubin (1977), in the E step, we mainly calculate the conditional expectation of certain sufficient statistics for all observations with the censored observations (incomplete data, data below LOQ) replaced by the conditional expectation using the parameters β_j and σ^2 which were obtained in the last iteration of the M step. Then the new parameters can be calculated from the expectation of sufficient statistics in this iteration M step.

Here is an example which shows the EM algorithm in detail. At the $(r+1)^{th}$ iteration, the E step should be given observed data (y_i) and parameters (β_j and σ^2) estimations from r^{th} iteration of the M step to calculate conditional expectation of sufficient statistics. This dissertation used $\sum_{i=1}^{n} x_{ij}y_i$ and $\sum_{i=1}^{n} y_i^2$ as sufficient statistics for the parameters in the complete data (Aitkin 1981; Thompson & Nelson 2003).

$$E\left(\sum_{i=1}^{n} x_{ij}y_i\right) = \sum_{i=1}^{n} \tau_i x_{ij}y_i + (1-\tau_i)E(y_i|y_i < a, \boldsymbol{\beta}, \sigma^2)) \qquad (2.17)$$

$$E\left(\sum_{i=1}^{n} y_i^2\right) = \sum_{i=1}^{n} \tau_i y_i^2 + (1-\tau_i)E(y_i^2|y_i < a, \boldsymbol{\beta}, \sigma^2)) \qquad (2.18)$$

where,

$$E(y_i \mid y_i < a, \boldsymbol{\beta}, \sigma^2) = \sum_{j=0}^{k} \beta_j x_{ij} - \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$$

$$E(y_i^2 \mid y_i < a, \boldsymbol{\beta}, \sigma^2) = var(y_i^2) + E^2(y_i)$$

$$= \sigma^2 + \left(\sum_{j=0}^{k} \beta_j x_{ij} - \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)\right)^2$$

$$= \sigma^2 + \left(\sum_{j=0}^{k} \beta_j x_{ij}\right)^2 + \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)^2 - 2\sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right) \cdot \sum_{j=0}^{k} \beta_j x_{ij}$$

$$= \sigma^2 + \left(\sum_{j=0}^{k} \beta_j x_{ij}\right)^2 + \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right) \cdot \left(\sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right) - 2\sum_{j=0}^{k} \beta_j x_{ij}\right)$$

$$= \sigma^2 + \left(\sum_{j=0}^{k} \beta_j x_{ij}\right)^2 + \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right) \cdot \left(w_i + \sum_{j=0}^{k} \beta_j x_{ij}\right)$$

with $w_i = \sum_{j=0}^{k} \beta_j x_{ij} - \sigma S\left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$.

Then, in the $(r+1)^{\text{th}}$ M step, new estimate $\boldsymbol{\beta}^{(r+1)}$ can be estimated by replacing the censored data w_i in the complete data set y_i by $\sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij} - \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij}}{\sigma_{(r)}}\right)$. And the new estimate $\sigma^2_{(r+1)}$ can be estimated by replacing two parts. First, as above, we use $\sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij} - \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij}}{\sigma_{(r)}}\right)$ to replace censored data w_i in complete data set y_i.

Second, $\sigma_{(r)}^2 + \left(\sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij}\right)^2 + \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij}}{\sigma_{(r)}}\right) \cdot \left(w_i + \sum_{j=0}^{k} \boldsymbol{\beta}^{(r)}{}_j x_{ij}\right)$ is used to replace the censored part in y_i^2.

In sum, $n\sigma_{(r+1)}^2 = \sum_{i=1}^{n} \tau_i \left(y_i - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 + (1 - \tau_i)\left(y_i - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2$

with

$$\left(y_i - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 = y_i^2 + \left(\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 - 2y_i \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}$$

$$= \sigma_{(r)}^2 + \left(\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 - \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right) \cdot \left(w_i + \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)$$

$$+ \left(\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 - 2\left(\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right)^2 + 2\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\, \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right)$$

$$= \sigma_{(r)}^2 - w_i \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right) - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\, \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right)$$

$$+2\sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\, \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right)$$

$$= \sigma_{(r)}^2 - \left(w_i - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}\right) \sigma_{(r)} S\left(\frac{a - \sum_{j=0}^{k} \beta_j^{(r)} x_{ij}}{\sigma_{(r)}}\right)$$

$$= \sigma_{(r)}^2 (1 - z_i^{(r)} S\left(z_i^{(r)}\right))$$

where $z_i = \left(\frac{a - \sum_{j=0}^{k} \beta_j x_{ij}}{\sigma}\right)$.

By the end, the iteration algorithm will stop when $\sigma^2_{(r)} = \sigma^2_{(r+1)}$. Then the MLE will be $\widehat{\sigma^2} = \sigma^2_{(r)} = \sigma^2_{(r+1)}$.

Chapter 3 - Tolerance Limit

A tolerance limit is the one-sided boundary of a tolerance interval. In drug withdrawal period studies, we focus on the one-sided upper boundary of the tolerance interval, which is the upper tolerance limit. I will introduce the classical approach to obtain tolerance limit, the FDA tolerance limit and the EMA tolerance limit in detail in this chapter. Moreover, a generalized variable approach with generalized pivotal quantities method is conducted to obtain tolerance limit. By the end, the generalized tolerance limit based on left censored data regression will be developed.

Suppose $F(x)$ is the cumulative distribution function (CDF) of a continuous random variable X, $F_X(x) = P(X \leq x)$. The inverse CDF can be written as the following function (3.1) by giving p $(0 < p < 1)$, the inverse CDF

$$F_X^{-1}(p) = \inf\{x: F_X(x) \geq p\} \qquad (3.1)$$

where, p is the proportion of the population with the CDF $F(x)$ which is less or equal to q_p,

q_p indicates the p quantile,

$F_X^{-1}(p)$ is the value of x which $F_X(x) = P(X \leq x) = p$.

Now, we can assume $X_1, X_2, X_3, \cdots\cdots, X_n$ to be a random sample from a CDF $F_X(x)$. Then $\boldsymbol{X}$ can be denoted as $\boldsymbol{X} = (X_1, X_2, X_3, \cdots\cdots, X_n)$. The one-sided tolerance interval for p content and $(1-\alpha)$ confidence is

$$P_{\boldsymbol{X}}\{P_x(X \leq I(\boldsymbol{X})|\boldsymbol{X}) \geq p\} = 1 - \alpha \qquad (3.2)$$

where, $0 \leq p \leq 1$, $0 \leq \alpha \leq 1$, and α is the significant level. Here, $I(\boldsymbol{X})$ shows that at least p proportion is less or equal to $I(\boldsymbol{X})$ with the confident level $(1-\alpha)$. Then $(-\infty, I(\boldsymbol{X})]$ is the one-sided tolerance interval and $I(\boldsymbol{X})$ is the upper tolerance limit. Since q_p is the p quantile of a distribution, we can know the above function (3.2) is able to be written as

$$P_{\boldsymbol{X}}\{q_p \leq I(\boldsymbol{X})\} = 1-\alpha \qquad (3.3)$$

Based on function (3.2), there is one important property (Krishnamoorthy & Mathew 2009). Let X follow a normal distribution with zero mean and variance σ^2 , $X \sim N(0, \sigma^2)$. It is independent to $Q \sim \frac{\chi_m^2}{m}$, where χ_m^2 is a chi square random variable with degree of freedom m. Suppose that $0 \leq p \leq 1$, $0 \leq \gamma \leq 1$, and Φ is the standard normal distribution function, then we can have a constant factor k satisfying

$$P_{X,Q}\left(\Phi\left(X + k\sqrt{Q}\right) \geq p\right) = \gamma \qquad (3.4)$$

where k is given by,

$$k = \sqrt{\sigma^2} \times t_{m;\gamma}\left(\frac{\mathcal{Z}_p}{\sqrt{\sigma^2}}\right) \qquad (3.5)$$

Here, $\mathcal{Z}_p$ is the p quantile of standard normal distribution Φ, $t_{m;\gamma}\left(\frac{\mathcal{Z}_p}{\sqrt{\sigma^2}}\right)$ is the γ quantile of the noncentral t distribution with degree of freedom m and noncentral parameter $\frac{\mathcal{Z}_p}{\sqrt{\sigma^2}}$.

This property can be proven by the following approach. We can see that the inner probability inequality part $\Phi\left(X + k\sqrt{Q}\right) \geq p$ in function (3.4) will hold if and only if $X + k\sqrt{Q} \geq \mathcal{Z}_p$. So, we can rewrite function (3.4) as

$$P_{X,Q}\left(X + k\sqrt{Q} \geq \mathcal{Z}_p\right) = P_{X,Q}\left(\frac{X - \mathcal{Z}_p}{\sqrt{Q}} \geq -k\right)$$

$$= P_{X,Q}\left(\sqrt{c}\frac{-X/\sqrt{c} + \mathcal{Z}_p/\sqrt{c}}{\sqrt{Q}} \leq k\right)$$

$$= P_{X,Q}\left(\sqrt{c}\frac{X/\sqrt{c} + \mathcal{Z}_p/\sqrt{c}}{\sqrt{Q}} \leq k\right)$$

$$= \gamma \qquad (3.6)$$

where c is the variance σ^2.

Note that the X and $-X$ are identically distributed, so that $P_{X,Q}\left(\sqrt{c}\frac{-X/\sqrt{c}+\mathcal{Z}_p/\sqrt{c}}{\sqrt{Q}} \leq k\right) = P_{X,Q}\left(\sqrt{c}\frac{X/\sqrt{c}+\mathcal{Z}_p/\sqrt{c}}{\sqrt{Q}} \leq k\right)$ in function (3.6). Since $X \sim N(0, \sigma^2)$, $X/\sqrt{c}$ follows a standard normal distribution, i.e. $X/\sqrt{c} \sim N(0,1)$, $c = \sigma^2$. In addition, $X/\sqrt{c} \sim N(0,1)$ and $Q \sim \frac{\chi_m^2}{m}$ are also independent to each other. Then, we can have

$$\frac{X/\sqrt{c} + \mathcal{Z}_p/\sqrt{c}}{\sqrt{Q}} \sim t_m\left(\frac{\mathcal{Z}_p}{\sqrt{c}}\right)$$

where $t_m\left(\frac{\mathcal{Z}_p}{\sqrt{c}}\right)$ is the noncentral *t* random variable with non-centrality parameter $\frac{\mathcal{Z}_p}{\sqrt{c}}$ and degree of freedom *m*. Hence, factor k satisfies function (3.4), $P_{X,Q}\left(\Phi\left(X + k\sqrt{Q}\right) \geq p\right) = \gamma$.

For the normal population, let's assume $X_1, X_2, X_3, \cdots\cdots, X_n$ be a sample from a normal distribution $N(\mu, \sigma^2)$. Here mean μ and variance σ^2 are unknown. Then, the sample mean $\bar{X}$ and sample variance S^2 can be written (De Gryze, Langhans & Vandebroek 2007) as

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Then the p quantile of the normal distribution $N(\mu, \sigma^2)$ is

$$q_p = \mu + \mathcal{Z}_p\sigma$$

where $\mathcal{Z}_p$ is the p quantile of a standard normal distribution $N(0,1)$. Note the upper confidence limit with $1-\alpha$ confidence level for q_p is the one-sided upper tolerance limit for a normal distribution with $(p, 1-\alpha)$. In general, the upper limit for q_p can be obtained by $p > 0.5$ (Krishnamoorthy & Mathew, 2009).

3.1 Classical Approach

Described in this section is the classic approach to calculate one-sided upper tolerance limits in a normal distribution setting. First, let's assume that the $(p, 1-\alpha)$ one-sided upper tolerance limit can be written as the form of $\bar{X} + kS$, k is the constant factor which we explained in the previous paragraphs. The one-sided upper tolerance limit is to determine that at least a proportion p of the population are less or equal than $\bar{X} + kS$ with confidence level $1-\alpha$. The following function (3.7) shows what we just described here.

$$P_{\bar{X},S}\{P(X < \bar{X} + kS|\bar{X}, S) > p\} = 1-\alpha \qquad (3.7)$$

where $X \sim N(\mu, \sigma^2)$. In addition, suppose $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$, $Z_n = \frac{\bar{X}-\mu}{\sigma} \sim N(0, \frac{1}{n})$, and $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$. $\chi^2_{n-1}$is the chi-square random variable with degree of freedom $n-1$.

Then the function (3.7) can be rewritten as

$$\begin{aligned} & P_{Z_n,U}\{P(Z < Z_n + kU|Z_n, U) > p\} \\ =\ & P_{Z_n,U}\{\Phi(Z_n + kU) > p\} \\ = 1-\alpha & \qquad (3.8) \end{aligned}$$

Since $Z_n = \frac{\bar{X}-\mu}{\sigma} \sim N(0, \frac{1}{n})$ and $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$ are independent to each other, based on function (3.4) and (3.5), we can have a constant factor k, which

$$k = \frac{1}{\sqrt{n}} t_{n-1;1-\alpha}\left(Z_p\sqrt{n}\right) \qquad (3.9)$$

where $\sigma^2 = \frac{1}{n}$, $\gamma = 1 - \alpha$, and $m = n - 1$. Here the $t_{n-1;1-\alpha}(\mathcal{Z}_p\sqrt{n})$ indicates the $1 - \alpha$ quantile of a non-central t distribution with degree of freedom $n - 1$ and the non-centrality parameter $\mathcal{Z}_p\sqrt{n}$.

So, we can have the $(p, 1 - \alpha)$ one-sided upper tolerance limit

$$\bar{X} + kS = \bar{X} + t_{n-1;1-\alpha}(\mathcal{Z}_p\sqrt{n})\frac{S}{\sqrt{n}}.$$

Actually, we can extend the $(p, 1 - \alpha)$ one-sided upper tolerance limit from normal population to linear regression. Y_i is the response variable for the i^{th} observation. x_i represents the covariates. Under normality assumption, the linear regression model can be represented as

$$Y = X\beta + \varepsilon$$

where, ε is the residual with $\varepsilon \sim N(0, \sigma^2)$, $Y = (Y_1, Y_2, Y_3, \cdots\cdots, Y_n)$ is the $n \times 1$ vector of the observations and X is $n \times m$ matrix with i^{th} row observation $x_i{}'$. Then we can know the mean of Y_i is $x'\beta$. Let $\hat{\beta}$ be the least square estimator of β, $\hat{\beta} = (X'X^{-1}X'Y)$. S^2 is the residual mean square with $S^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n - m}$.

Now, let $k(x)$ be the tolerance factor which can be calculated to have the one side upper tolerance interval with the form $(-\infty, x'\hat{\beta} + k(x)S]$. Here $x'\hat{\beta} + k(x)S$ is the upper tolerance limit. Based on the theory we just shown above, given $\hat{\beta}$ and S^2 the content of the one side upper tolerance interval for linear regression is

$$C(x; \hat{\beta}, S) = P_{Y(x)}(Y(x) \leq x'\hat{\beta} + k(x)S \,|\, \hat{\beta}, S) \qquad (3.10)$$

Then, the constant tolerance factor $k(x)$ should satisfy

$$P_{\hat{\beta},S}\big(\, C(x;\hat{\beta},S) \geq P\big) = 1-\alpha \qquad (3.11)$$

Note that $\hat{\beta} = (X'X^{-1}X'Y)$, $Y \sim N(X\beta, \sigma^2 I)$. Then expectation of $\hat{\beta}$ is

$E(\hat{\beta}) = E((X'X)^{-1}X'X\beta) = \beta$. The variance of parameter β can be written as $var(\hat{\beta}) = (X'X)^{-1}X'\sigma^2 I[(X'X)^{-1}X']' = \sigma^2(X'X)^{-1}$.

So we can have $Z = \frac{Y(x)-x'\widehat{\beta}}{\sigma} \sim N(0,1)$, $Z_x = \frac{\widehat{\beta}-\beta}{\sigma} \sim N(0,(X'X)^{-1})$, $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-m}}{n-m}$ and all random variables are independent. χ^2_{n-m} is the chi-square random variable with degree if freedom $n-m$.

Then we can rewrite function (3.10) as

$$C(x;\hat{\beta},S) = P_z(z \leq x'Z_x + k(x)U \mid Z_x, U) \qquad (3.12)$$

Note that $var(x'Z_x) = x'var(Z_x)x = x'(X'X)^{-1}x$. Hence, $x'Z_x$ follows normal distribution which $x'Z_x \sim N(0, x'(X'X)^{-1}x)$.

Let d^2 be the variance of $x'Z_x$, $d^2 = x'(X'X)^{-1}x$, and $v = \frac{x'Z_x}{d}$. So v is a standard normal distribution, $v \sim N(0,1)$. Now, function (3.10) can be rewritten as

$$C(x;\hat{\beta},S) = \Phi(dv + k(d)U) = C(d;v,U) \qquad (3.13)$$

where, Φ is the cumulative density function (cdf) of a standard normal distribution. $k(d)$ is the new notation instead of $k(x)$ since d is a function of x. Based on function (3.13) and (3.11), the constant tolerance factor $k(d)$ will satisfy

$$P_{v,U}(\Phi(dv + k(d)U) \geq P) = 1-\alpha$$

Let $X = dv$ and $Q = U^2 = \frac{S^2}{\sigma^2}$, so that $X \sim N(0, d^2)$ and Q$\sim \frac{\chi^2_{n-m}}{n-m}$. With function (3.5), we can know the constant tolerance factor $k(d)$ can be calculated by

$$k(d) = d \times t_{n-m;1-\alpha}(\frac{Z_p}{d})$$

where $t_{n-m;1-\alpha}(\frac{Z_p}{d})$ is the $1-\alpha$ quantile of a noncentral t distribution with degree of freedom $n-m$ and the non-centrality parameter $\frac{Z_p}{d}$.

In the simple linear regression, we can have

$$d^2 = \frac{1}{n} + c^2 \qquad (3.14)$$

where $c^2 = \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. It can be proved by the following steps. Let $x = \begin{pmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_n \end{pmatrix}$, so that $x' = \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$. Then we can have

$$x'x = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$(x'x)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$x'(x'x)^{-1}x = \frac{\sum_{i=1}^{n} x_i^2 - 2x\sum_{i=1}^{n} x_i + nx^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

In addition, we can know

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2+n\bar{x}^2-2\sum_{i=1}^{n}x_i\bar{x}$$

$$=\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}$$

and

$$(x-\bar{x})^2=x^2+\bar{x}^2-2x\bar{x}$$

$$=x^2+(\frac{\sum_{i=1}^{n}x_i}{n})^2-2x\frac{\sum_{i=1}^{n}x_i}{n}.$$

Then

$$x'(x'x)^{-1}x=\frac{n\left(x^2-2\frac{x}{n}\sum_{i=1}^{n}x_i+\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2-\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2+\frac{\sum_{i=1}^{n}x_i^2}{n}\right)}{n(\sum_{i=1}^{n}(x_i-\bar{x})^2)}$$

$$=\frac{n\left((x-\bar{x})^2-\left(\frac{\sum_{i=1}^{n}x_i}{n}\right)^2+\frac{\sum_{i=1}^{n}x_i^2}{n}\right)}{n(\sum_{i=1}^{n}(x_i-\bar{x})^2)}$$

$$=\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{\sum_{i=1}^{n}x_i^2-\frac{(\sum_{i=1}^{n}x_i)^2}{n}}{n(\sum_{i=1}^{n}(x_i-\bar{x})^2)}$$

$$=\frac{1}{n}+c^2$$

so that we have proven $d^2=\frac{1}{n}+c^2$, where $c^2=\frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ for the simple linear regression.

3.2 FDA and EMA Tolerance Limit

The FDA's CVM suggests 99% setting for the proportion p with 95% confident level. In other words, the FDA requires the tolerance limits to cover 99% of the population with 95% confidence. However, the EMA guidance suggests a 95% population coverage with 95% confidence. The proportion p is set as 0.95 instead of 0.99. In general, the FDA tolerance interval will have bigger range than EMA approach. It may lead to more conservative withdrawal period.

The FDA (2005) specified guidance on how to calculate the tolerance limit, which is shown below.

$$Y = \beta_0 + \beta_1 x + k(d) S_{xy}$$

where, $k(d) = d \times t_{n-m;1-\alpha}(\frac{Z_p}{d})$,

$d^2 = \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, and

S_{xy} is the residual error, Z_p is 99[th] percentile of standard normal distribution with $p = 0.99$, $1-\alpha$ is 95% confidence level with $\alpha = 0.05$, n is sample size and $m = 2$.

It is clear that the FDA tolerance limit is as same as we described in last section. In addition to the different setting on proportion p, where $p = 0.95$, the EMA tolerance limit is based on Kurt (1971) tolerance limit instead of the approach we just described before. However, the two approach are very similar. The results are close to each other when they are based on same proportion p and confidence level $1-\alpha$ (EMA-CVMP 2016). The function below shows the EMA tolerance limit in detail.

$$Y = \beta_0 + \beta_1 x + k_t S_{xy}$$

where, $k_t = \frac{\sqrt{(2n-4)}}{(2n-4)-Z_{1-\alpha}^2}(\sqrt{(2n-4)}Z_{1-\gamma} + Z_{1-\alpha}W_n$

$$W_n = \sqrt{Z_{1-\gamma}^2 + ((2n-4) - Z_{1-\alpha}^2)(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}})}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

S_{xy} is the residual error. $Z_{1-\gamma}$ and $Z_{1-\alpha}$ are $(1-\gamma)$ and $(1-\alpha)$ percentile of standard normal distribution. Here $(1-\gamma) = (1-\alpha) = 0.95$ based on EMA guidance.

3.3 Generalized Variable Approach

Beside the classical approach, there are additional procedures to obtain the one-sided tolerance limits. In fact, for a normal distribution $N(\mu, \sigma^2)$, the computation of the $(p, 1-\alpha)$ one-sided upper tolerance limit is equivalent to the computation of the $1-\alpha$ upper confidence limit for the p quantile of the normal distribution (Krishnamoorthy & Mathew 2009), $q_p = \mu + Z_p\sigma$. The modified large sample (MLS) confidence interval approach was proposed by Graybill and Wang (1980) and studied by Burdick and Graybill (1992). Unfortunately, the large sample method does not perform well for small sample sizes which is often the situation in practice (Borror, Montgomery & Runger, 1997). An alternative method is called the generalized variable approach. It contains two important concepts: the generalized p-value approach for hypothesis testing, which was introduced by Tsui and Weerahandi (1989) and the generalized confidence interval (Weerahandi, 1995).

Based on the generalized p-value and the generalized confidence interval, the generalized variable approach (generalized inference procedure) is founded. The generalized variable approach has been proven to be extremely useful to obtain confidence intervals for complex situations when the standard or classic approaches are hard to apply (Krishnamoorthy & Mathew, 2009). In order to conduct the generalized confidence interval, we will introduce the generalized pivotal quantity approach (GPQ).

Let X be a random sample from a distribution $F_X(x;\ \theta, \delta)$. θ is the scalar parameter we are interested in. δ is the other nuisance parameter. x is an observation, which represents the data. The generalized confidence interval for interest parameter θ can be obtained by using the percentile of the GPQ, such as $G(X; x, \theta)$. Here X, x, and θ satisfy two conditions.

First, given x, the distribution of $G(X; x, \theta)$ is free of all unknown parameters. Second, the observed value of the distribution $G(X; x, \theta)$ is θ, which is the parameter we are interested in, when $X = x$. Note that since the interest parameter θ is an unknown parameter, it is not observable. When both conditions hold, the $1 - \alpha$ confidence interval for θ is the p quantile of $G(X; x, \theta)$. For example, $(G_{\frac{\alpha}{2}},\ G_{1-\frac{\alpha}{2}})$ is the $1 - \alpha$ confidence interval for θ when G_p is the p quantile of $G(X; x, \theta)$. Then, $(G_{\frac{\alpha}{2}},\ G_{1-\frac{\alpha}{2}})$ is the generalized confidence interval.

For the location scale family, GPQ has a useful property. A continuous distribution belongs to location scale family if its probability density function (pdf) can be written as in the following form

$$f(x;\ \mu, \sigma) = \frac{1}{\sigma} g\left(\frac{x-\mu}{\sigma}\right) \qquad (3.15)$$

where $-\infty \leq x \leq \infty$, $-\infty \leq \mu \leq \infty$ and $\sigma \geq 0$.

The function $g(\cdot)$ is a completely specified pdf. The terms μ and σ are location and scale parameters for the pdf. For example, the normal distribution is a location scale family since we can write the normal distribution pdf in the form

$$f(x;\ \mu, \sigma) = \frac{1}{\sigma} \phi(\frac{x-\mu}{\sigma})$$

where $\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

Now, let $X_1, X_2, X_3, \cdots\cdots, X_n$ be a sample from a certain distribution with location and scale parameter μ and σ. The estimators $\hat{\mu}(X_1, X_2, X_3, \cdots\cdots, X_n)$ of μ and $\hat{\sigma}(X_1, X_2, X_3, \cdots\cdots, X_n)$ of σ are equivariant if for any constant number *a* and *b*, *a > 0*,

$$\hat{\mu}(aX_1 + b, aX_2 + b, \cdots\cdots, aX_n + b) = a\hat{\mu}(X_1, X_2, \cdots\cdots, X_n) + b$$

$$\hat{\sigma}(aX_1 + b, aX_2 + b, \cdots\cdots, aX_n + b) = a\hat{\sigma}(X_1, X_2, \cdots\cdots, X_n)$$

This indicates that the sample mean $\bar{X}$ and sample variance S^2 are equivariant estimators for a normal mean and variance. If the sample is from a location family, $X_1, X_2, X_3, \cdots\cdots, X_n$ is from a continuous distribution with pdf in the form of function (3.15), and $\hat{\mu}(X_1, X_2, X_3, \cdots\cdots, X_n)$ amd $\hat{\sigma}(X_1, X_2, X_3, \cdots\cdots, X_n)$ are equivariant estimators of μ and σ, then $\frac{\hat{\mu}-\mu}{\sigma}$, $\frac{\hat{\sigma}}{\sigma}$ and $\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ are all pivotal quantities. That means their distributions do not depend on any parameters.

We can prove this by assuming $Z_i = \frac{X_i-\mu}{\sigma}$, $i = 1,2,3, \cdots\cdots, n$. The sample $X_1, X_2, X_3, \cdots\cdots, X_n$ is from a location scale distribution family. The joint distribution of $Z_i = \frac{X_i-\mu}{\sigma}$ are free of unknown parameters. Because $\hat{\mu}(X_1, X_2, X_3, \cdots\cdots, X_n)$ and $\hat{\sigma}(X_1, X_2, X_3, \cdots\cdots, X_n)$ are equivariant, we can have

$$\frac{\hat{\mu}(X_1, X_2, X_3, \cdots\cdots, X_n) - \mu}{\sigma} = \hat{\mu}\left(\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \cdots\cdots, \frac{X_n - \mu}{\sigma}\right)$$

$$= \hat{\mu}(Z_1, Z_2, Z_3, \cdots\cdots, Z_n)$$

$$\frac{\hat{\sigma}(X_1, X_2, X_3, \cdots\cdots, X_n) - \mu}{\sigma} = \hat{\sigma}\left(\frac{X_1 - \mu}{\sigma}, \frac{X_2 - \mu}{\sigma}, \cdots\cdots, \frac{X_n - \mu}{\sigma}\right)$$

$$= \hat{\sigma}(Z_1, Z_2, Z_3, \cdots\cdots, Z_n)$$

Then we know $\frac{\hat{\mu}-\mu}{\sigma}$ and $\frac{\hat{\sigma}}{\sigma}$ are pivotal quantities. Since $\frac{\hat{\mu}-\mu}{\hat{\sigma}} = \frac{\hat{\mu}-\mu}{\sigma} \times \frac{\hat{\sigma}}{\sigma}$, $\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ is also a pivotal quantity. Then the GPQ of interest location and scale parameters μ and σ can be developed as the following functions based on pivotal quantities. Assume $\widehat{\mu_0}$ and $\widehat{\sigma_0}$ are observed values of the equivariant estimators $\hat{\mu}$ and $\hat{\sigma}$. Then the value of GPQ for μ should be μ when $(\hat{\mu}, \hat{\sigma}) = (\widehat{\mu_0}, \widehat{\sigma_0})$ based on the second condition for GPQ we mentioned before.

Hence, the GPQ for μ can be written as

$$G_\mu(\hat{\mu}, \hat{\sigma}; \widehat{\mu_0}, \widehat{\sigma_0}) = \widehat{\mu_0} - \left(\frac{\hat{\mu} - \mu}{\hat{\sigma}}\right)\widehat{\sigma_0} \qquad (3.16)$$

Here, the value of function (3.16) is μ at $(\hat{\mu}, \hat{\sigma}) = (\widehat{\mu_0}, \widehat{\sigma_0})$. Keeping in mind, since $\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ is a pivotal quantity, given $\widehat{\mu_0}$ and $\widehat{\sigma_0}$, the distribution of G_μ does not depend on any unknown parameters. In addition, the GPQ for scale parameter σ^2 can be written as

$$G_{\sigma^2}\left(\hat{\sigma}^2; \widehat{\sigma_0}^2\right) = \frac{\sigma^2}{\hat{\sigma}^2}\widehat{\sigma_0}^2 \qquad (3.17)$$

Similar as G_μ, G_{σ^2} does not depend on any parameters since $\frac{\sigma^2}{\hat{\sigma}^2}$ is pivotal quantity. Both $G_\mu(\hat{\mu}, \hat{\sigma}; \widehat{\mu_0}, \widehat{\sigma_0})$ and $G_{\sigma^2}\left(\hat{\sigma}^2; \widehat{\sigma_0}^2\right)$ satisfy the two conditions for GPQ. So, $G_\mu(\hat{\mu}, \hat{\sigma}; \widehat{\mu_0}, \widehat{\sigma_0})$ and $G_{\sigma^2}\left(\hat{\sigma}^2; \widehat{\sigma_0}^2\right)$ are GPQ for location parameter μ and scale parameter σ^2.

GPQ has another great property that the normal and non-normal parameters of the GPQ for any function of (μ, σ^2) can be easily obtained by substitution of the function for (G_μ, G_{σ^2}). For example, if we want to conduct any inference for the function of $f(\mu, \sigma^2)$, we can then have GPQ by calculating $f(G_\mu, G_{\sigma^2})$ (Krishnamoorthy & Mathew 2009). This property can easily

help us make inference on normal and non-normal parameters in the future. Let's see some examples. Suppose $X_1, X_2, X_3, \cdots\cdots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$. We can define the sample mean $\bar{X}$ and sample variance S^2 to be

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Let $\bar{x}$ and s^2 be the observed value of $\bar{X}$ and S^2. Then we can have $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = S^2$, $\widehat{\mu_0} = \bar{x}$ and $\widehat{\sigma_0}^2 = s^2$. Based on function (3.16), the GPQ for the mean can be written as

$$\begin{aligned} G_\mu = G_\mu(\hat{\mu}, \hat{\sigma}; \widehat{\mu_0}\,, \widehat{\sigma_0}) &= \widehat{\mu_0} - \left(\frac{\hat{\mu} - \mu}{\hat{\sigma}}\right)\widehat{\sigma_0} \\ &= \widehat{\mu_0} - \frac{(\bar{X} - \mu)}{S}s \\ &= \widehat{\mu_0} - \frac{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}}{\frac{S}{\sigma}}\frac{s}{\sqrt{n}} \\ &= \widehat{\mu_0} - \frac{Z}{U}\frac{s}{\sqrt{n}} \\ &= \bar{x} + \frac{Z}{U}\frac{s}{\sqrt{n}} \qquad (3.18) \end{aligned}$$

where $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, which follows standard normal distribution, $Z \sim N(0,1)$. Also Z and $-Z$ have same distribution. $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$. Z and U^2 are independent of each other. Since $\frac{Z}{U} \sim t_{n-1}$, the function (3.18) can be written as

$$G_\mu = G_\mu(\bar{X}, S; \bar{x}, s) = \bar{x} + t_{n-1}\frac{s}{\sqrt{n}}$$

Then the generalized confidence interval with significance level α is ($\bar{x} + t_{n-1;\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$, $\bar{x} + t_{n-1;1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$). It is clear to see this generalized confidence interval is as same as a common t interval. Similar to GPQ for normal mean, based on function (3.17), the GPQ for normal variance can be written as

$$G_{\sigma^2}\left(\hat{\sigma}^2;\ \widehat{\sigma_0}^2\right) = \frac{\sigma^2}{\hat{\sigma}^2}\widehat{\sigma_0}^2 = \frac{\sigma^2}{S^2}s^2 = \frac{s^2}{U^2} \qquad (3.19)$$

where $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$. Then the generalized confidence interval for σ^2 with significant level α is ($\frac{(n-1)s^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}$, $\frac{(n-1)s^2}{\chi^2_{n-1;\frac{\alpha}{2}}}$), which is the regular interval for σ^2.

We can see that the generalized variable approach can help us get the exact inference for normal parameters. The solutions are reduced to usual intervals. Because of this property, we mentioned that the GPQ for any function of $(\mu\ , \sigma^2)$ can be easily obtained by substation of the function for $(G_\mu\ , G_{\sigma^2})$. For example, let's obtain the lognormal mean. Note that if Y follows lognormal distribution then $X = \ln(Y)$ is a normal distribution with parameters μ and σ^2. The lognormal mean is $E(Y) = \exp(\ \mu\ + \frac{\sigma^2}{2})$. Then the GPQ for the lognormal mean $E(Y)$ can be obtained by GPQ for parameters μ and σ^2, with $\tau = (\ \mu\ + \frac{\sigma^2}{2})$.

Let $\bar{X}$ and S^2 be the mean and variance of the normal distribution which are obtained by log transformation from the log normal distribution. The GPQ for τ should be

$$G_\tau = G_\mu + \frac{G_{\sigma^2}}{2} = \bar{x} + \frac{Z}{U}\frac{s}{\sqrt{n}} + \frac{(n-1)s^2}{2U^2}$$

which is the GPQ for the function of (G_μ , G_{σ^2}). This property can help in future studies when we assume the distribution is not normal or want to have inference on $f(G_\mu , G_{\sigma^2})$.

Compared to the classical approach, the generalized variable approach one-sided upper tolerance interval for normal population is quite simple. Since the computation of the $(p, 1-\alpha)$ one-sided upper tolerance limit for the normal distribution $N(\mu, \sigma^2)$ is equivalent to the computation of the $1-\alpha$ upper confidence limit for the p quantile of the normal distribution, $q_p = \mu + \mathcal{Z}_p\sigma$ (Krishnamoorthy & Mathew, 2009), we can generate the generalized variable approach using the following steps.

Based on the GPQ property we mentioned before, the GPQ for $q_p = \mu + \mathcal{Z}_p\sigma$ can be written as

$$
\begin{aligned}
G_{q_p} &= G_\mu + \mathcal{Z}_p\sqrt{G_{\sigma^2}} \\
&= \bar{x} + \frac{Z}{U}\frac{s}{\sqrt{n}} + \mathcal{Z}_p\sqrt{\frac{s^2}{U^2}} \\
&= \bar{x} + \frac{Z}{U}\frac{s}{\sqrt{n}} + \mathcal{Z}_p\frac{s}{U} \\
&= \bar{x} + \frac{Z + \mathcal{Z}_p\sqrt{n}}{U}\frac{s}{\sqrt{n}} \\
&= \bar{x} + \frac{1}{\sqrt{n}}t_{n-1}(\mathcal{Z}_p\sqrt{n})s
\end{aligned}
$$

where $Z = \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}$, which follows standard normal distribution, i.e. $Z \sim N(0,1)$. $U^2 = \frac{S^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$ with Z and U^2 independent. G_μ and G_{σ^2} are from functions (3.18) and (3.19) respectively. Hence, the distribution G_{q_p} does not depend on any unknown parameters. The percentile of G_{q_p} is the

confidence limit for $q_p = \mu + Z_p\sigma$. For example, the one-sided upper tolerance interval with generalized variable approach can be obtained with the $1 - \alpha$ quantile of the G_{q_p} by $\bar{x} + t_{n-1;1-\alpha}(Z_p\sqrt{n})s$.

3.4 Generalized Pivotal Quantity for Left Censored Data Regression

For censored data, we can use pivotal quantities to make inference and the results seem to be satisfactory for small sample sizes (Schmee, Gladstein & Nelson, 1985). Here, we will develop the GPQ approach for the left censored data regressing setting to generate the $(p, 1-\alpha)$ one-sided upper tolerance limit. As we mentioned before, the $(p, 1-\alpha)$ one-sided upper tolerance limit is the $1-\alpha$ upper confidence limit for the p quantile of the distribution, $q_p = \mu + \mathcal{Z}_p\sigma$. Now, we need to construct the $1-\alpha$ upper confidence limit for the p quantile of the distribution, $q_p = \mu + \mathcal{Z}_p\sigma$. Here, the $\mathcal{Z}_p$ is the p quantile of a standard normal distribution.

Let's assume $\widehat{\mu_0}$ and $\widehat{\sigma_0}$ are observed values from left censored data regression MLE of $(\hat{\mu}, \hat{\sigma})$. Note that $\hat{\mu} = X\hat{\beta}$ in the regression setting. Then the GPQ for left censored data regression is

$$
\begin{aligned}
G_{q_p} &= G_\mu + \mathcal{Z}_p G_\sigma \\
&= \widehat{\mu_0} - \frac{\hat{\mu}-\mu}{\hat{\sigma}}\widehat{\sigma_0} + \mathcal{Z}_p\frac{\sigma}{\hat{\sigma}}\widehat{\sigma_0} \\
&= \widehat{\mu_0} + \frac{\mathcal{Z}_p - (\hat{\mu}-\mu)/\sigma}{\hat{\sigma}/\sigma}\widehat{\sigma_0} \\
&= \widehat{\mu_0} + \frac{\mathcal{Z}_p - \hat{\mu}^*}{\hat{\sigma}^*}\widehat{\sigma_0} \\
&= \widehat{\mu_0} + \mathcal{Z}_p^*\widehat{\sigma_0}
\end{aligned}
$$

where $\hat{\mu}^* = \frac{\hat{\mu}-\mu}{\sigma}$, $\hat{\sigma}^* = \frac{\hat{\sigma}}{\sigma}$ and $\mathcal{Z}_p^* = \frac{\mathcal{Z}_p - \hat{\mu}^*}{\hat{\sigma}^*}$.

Now, let $\mathcal{Z}_{p;\alpha}^*$ be the α quantile of $\mathcal{Z}_p^*$. Then we can define the $(p, 1-\alpha)$ one side upper tolerance interval as

$$\widehat{\mu_0} + \mathcal{Z}^*_{p;1-\alpha}\,\widehat{\sigma_0} \qquad (3.20)$$

Right now, we can use resampling methods to estimate the percentile of $\mathcal{Z}^*_p$. The following paragraphs describe the resampling method in detail.

1. Since $\mathcal{Z}_p$ is the p quantile of a standard normal distribution, we generate a sample $\mathcal{Z}_1, \mathcal{Z}_2, \mathcal{Z}_3, \cdots\cdots, \mathcal{Z}_n$ from a standard normal distribution, $N(0,1)$. Then sort the sample in ascending order as $\mathcal{Z}_{(1)}, \mathcal{Z}_{(2)}, \mathcal{Z}_{(3)}, \cdots\cdots, \mathcal{Z}_{(n)}$. So, we can calculate the p^{th} quantile of a standard normal distribution $\mathcal{Z}_p$.
2. Calculate $\hat{\mu}^*$ and $\hat{\sigma}^*$ based on $\hat{\mu}^* = \frac{\hat{\mu}-\mu}{\sigma}$ and $\hat{\sigma}^* = \frac{\hat{\sigma}}{\sigma}$, where $\hat{\mu} = X\hat{\beta}$. $\hat{\mu}$ and $\hat{\sigma}$ are MLE from left censored data regression.
3. Obtain $\mathcal{Z}^*_p = \frac{\mathcal{Z}_p - \hat{\mu}^*}{\hat{\sigma}^*}$.
4. Repeat steps 1-3 for a large number of times. Here 10,000 repetitions were used.

Then, the $1-\alpha$ percentile of such 10,000 simulations of $\mathcal{Z}^*_p$ is the $\mathcal{Z}^*_{p;1-\alpha}$. Now, we can have the $(p, 1-\alpha)$ one-sided upper tolerance interval with $\widehat{\mu_0} + \mathcal{Z}^*_{p;1-\alpha}\,\widehat{\sigma_0}$.

Chapter 4 - Simulation and Results

To evaluate the performance of the proposed approaches, simulation studies are conducted in this chapter in two sections. One is for left censored data regression based on data below LOQ. The other is with respect to an upper tolerance limit using parameter estimates from a left censored data regression. In section 4.1, simulation studies will be shown to compare the model performance between left censored data regression and some existing methods with observations below the LOQ under different scenarios. This portion of work indicates that the proposed left censored data regression method provides results that are comparable to currently used approachs, such as the EMA's approach of assigning values below the LOQ to half the LOQ and the FDA's approach of omitting value below LOQ. The performance of left censored data regression with varying proportions of LOQ data in the total data is also studied. Through these analyses we can explore the advantages and disadvantages for left censored data regression. In section 4.2, a simulation study is conducted to evaluate generalized pivotal quantity tolerance limit performance and compare it to the FDA and EMA tolerance limits.

The guideline entitled "The General Principles for Evaluating the Safety of Compounds Used in Food-Producing Animals" (FDA 2005) claims that for the studies of residue in tissue, sufficient residue data are generally provided from the target tissue of 20 animals with five animals being slaughtered at each of four evenly spaced time points. The FDA also suggests that data below the LOQ should be excluded. Then the minimum sample size above the LOQ should be 20. The guideline "Approach Towards Harmonization of Withdrawal Periods" (EMA-CVMP 2016) recommends that, depending on the drug, the type of study and animal species, 4-10 animals in each time point are required. In addition, it should be kept in mind that 3 time points

are necessary to allow a meaningful linear regression analysis. In general, from a statistical standpoint, residue data from a minimum of 16 animals with four animals being euthanized at each of four appropriately distributed time intervals are recommended (EMA/CVMP/VICH/463199/2009). A sufficient number of birds should be used to obtain at least 6 samples at each slaughter time for tissue residue studies in poultry (EMA/CVMP/VICH/463199/2009).

With these considerations in mind, the lowest sample size was set at 24, which satisfies both regulatory guidelines. In this case, there will be 6 samples at each of the four evenly spaced time points. This should ensure an adequate sample size even after excluding data below LOQ. For example, when approximately 15% of the data is expected to be below LOQ, the expected amount of observations above LOQ is 20 in which case the recommended sample sized under both regulatory guidelines is obtained. The 4 evenly spaced time points studied here are at 6, 12, 18, and 24 days after dosing.

4.1 Left Censored Data Regression

In order to set up the simulation parameters in this section, a realistic true linear regression model must be specified to represent the drug residue depletion over time. Let $\sigma = 1$ and $\mu(t) = \beta_0 - \beta_1 t$, where $\beta_1 > 0$ and t is the variable which indicates time after dosing, assume that $Y(t) \sim N(\mu(t), 1)$. Then $Z = \frac{Y(t)-\mu(t)}{1} = Y(t) - \beta_0 + \beta_1 t$ should follow a standard normal distribution, i.e. $Z \sim N(0,1)$. Since there are 4 time points, each time point contains 25% of the total observations. In addition, it is possible for observations below LOQ to occur not only at the last sampling time point but also at earlier time points with a certain low probability. Here, we assume the probability that the observations below LOQ occur at the last time point is 0.95. It indicates that the probability which the observations below LOQ occur at the third time point is 0.05.

First, let us assume there are 5% of data below LOQ. At the last sampling time point *t* = 24, the LOQ data should be 20% (5%/25%) of samples. In addition, we assume the probability LOQ data occurs at this last sampling point is 0.9. Then the probability $P(Z \leq LOQ - \beta_0 - \beta_1 t) = P(Z \leq LOQ - \beta_0 - \beta_1 \times 24) = \frac{0.05 \times 0.9}{0.25} = 0.18$. We then obtain a z-score of $Z_{0.18} = -0.915$. At the third sampling time point *t* = 18, the probability of LOQ data occurring is 0.1. Then the probability $P(Z \leq LOQ - \beta_0 - \beta_1 t) = P(Z \leq LOQ - \beta_0 - \beta_1 \times 18) = \frac{0.05 \times 0.1}{0.25} = 0.02$. This results in a a z-score of $Z_{0.02} = -2.054$. Now there are two equations to solve:

$$LOQ - \beta_0 - \beta_1 \times 18 = -2.054$$

$$LOQ - \beta_0 - \beta_1 \times 24 = -0.915$$

For simplicity, assume LOQ = 1. Then, we obtain $\beta_1 = -0.2$, $\beta_0 = 6.7$. This means our true linear regression is

$$Y(t) = 6.7 - 0.2t$$

Next, we need to reset LOQ level to have 10% of data below LOQ, which

$$LOQ - 6.7 + 0.2 \times 24 = Z_{\frac{0.1\times 0.9}{0.25}} = -0.358.$$

Then, the LOQ is 1.542. Based on the same calculation process, the LOQ should be 2 with 15% of data below the LOQ, which $LOQ - 6.7 + 0.2 \times 24 = Z_{\frac{0.15\times 0.9}{0.25}} = 0.1.$

For each scenarios described above (5, 10 and 15% of data below the LOQ) the total sample sizes was set to 24 observations (6 samples at each of four evenly spaced time points), 32 observations (8 samples at each of four evenly spaced time points), 40 observations (10 samples at each of four evenly spaced time point.) and 48 observations (12 samples at each of four evenly spaced time points).

Three approaches of handling LOQ data are compared in this section. These are the newly proposed left censored data regression, the EMA's approach (assigning values below the LOQ to half the LOQ) and the FDA approach (omitting value below LOQ). After *n* = 1000 simulations, the bias and root mean square error (RMSE) of parameter estimations are computed and then reported in Table 4.1.

According to our simulation results, when sample size increases, the bias and RMSE of β_0 and β_1 both decrease over all three approaches. Left censored data regression performs very well when the sample size is 48, with the bias of β_0 smaller than 0.001. When the proportion of LOQ data increases, the bias and RMSE of β_0 and β_1 both increase. However, left censored data regression always has lowest RMSE for β_0 and β_1 of the three approaches. In addition, the

amount of increase in the RMSE of β_0 and β_1 as the proportion of LOQ data increases with left censored data regression is very small – much smaller than with the other two approaches. Based on the two results above it is clear that the left censored data regression has a more stable estimation of β_0 and β_1 than the others.

Regardless of simulation settings, the approach of omitting LOQ data preforms the worst with highest bias and RMSE for β_0 and β_1. This finding is as same as the conclusion given in EMA-CVMP (2016). The half LOQ method performs better than the omission approach in all simulation scenarios. However, the proposed approach, i.e. left censored data regression, performs the best. It resulted in the lowest bias and RMSE for β_1 and β_0 especially when the proportion of LOQ is large. The only exception is the bias of β_0 with 5% LOQ, where left censored regression has a larger bias than the half LOQ approach. However, the RMSE of β_0 with left censored data regression is still smaller than the half LOQ method.

In summary, the left censored data regression preforms the best of the three studied approaches. It has much smaller bias and RMSE for β_0 and β_1 than the half LOQ and omitting LOQ approaches, especially when large proportions of the data are below the LOQ.

Table 4.1 Simulation Results.

LCR is the proposed approach, left censored data regression. Half LOQ is the EMA's approach (assigning values below the LOQ to half the LOQ). The omit is FDA approach (omitting value below LOQ).

(a). Sample size = 24, $\beta_0 = 6.7, \beta_1 = -0.2$

		$\widehat{\beta_0}$		$\widehat{\beta_1}$	
		Bias	**RMSE**	**Bias**	**RMSE**
5%	LCR	0.008	0.414	-0.001	0.004
	Half LOQ	-0.004	0.426	0.001	0.004
	Omit	-0.147	0.517	0.015	0.021
10%	LCR	0.002	0.424	-0.001	0.005
	Half LOQ	0.015	0.437	-0.002	0.006
	Omit	-0.237	0.557	0.027	0.032
15%	LCR	-0.002	0.424	-0.001	0.005
	Half LOQ	0.041	0.457	-0.005	0.011
	Omit	-0.318	0.596	0.038	0.043

(b). Sample size = 32, $\beta_0 = 6.7, \beta_1 = -0.2$

		$\widehat{\beta_0}$		$\widehat{\beta_1}$	
		Bias	**RMSE**	**Bias**	**RMSE**
5%	LCR	0.008	0.359	-0.001	0.003
	Half LOQ	-0.002	0.369	0.001	0.004
	Omit	-0.144	0.454	0.015	0.020
10%	LCR	0.003	0.367	-0.001	0.005
	Half LOQ	0.017	0.378	-0.002	0.006
	Omit	-0.239	0.502	0.027	0.031
15%	LCR	-0.001	0.367	<0.001	0.001
	Half LOQ	0.043	0.399	-0.005	0.010
	Omit	-0.322	0.544	0.038	0.042

(c). Sample size = 40, $\beta_0 = 6.7, \beta_1 = -0.2$

		$\widehat{\beta_0}$		$\widehat{\beta_1}$	
		Bias	**RMSE**	**Bias**	**RMSE**
5%	LCR	0.006	0.323	-0.001	0.003
	Half LOQ	-0.002	0.332	0.001	0.004
	Omit	-0.145	0.414	0.015	0.019
10%	LCR	0.002	0.329	-0.001	0.004
	Half LOQ	0.017	0.340	-0.002	0.006
	Omit	-0.242	0.466	0.027	0.030
15%	LCR	-0.001	0.330	<0.001	0.001
	Half LOQ	0.043	0.357	-0.006	0.010
	Omit	-0.328	0.516	0.039	0.042

(d). Sample size = 48, $\beta_0 = 6.7, \beta_1 = -0.2$

		$\widehat{\beta_0}$		$\widehat{\beta_1}$	
		Bias	**RMSE**	**Bias**	**RMSE**
5%	LCR	0.001	0.296	<-0.001	0.002
	Half LOQ	-0.006	0.303	<-0.001	0.003
	Omit	-0.149	0.385	0.015	0.018
10%	LCR	-0.001	0.301	<-0.001	0.003
	Half LOQ	0.013	0.311	-0.002	0.005
	Omit	-0.247	0.440	0.027	0.030
15%	LCR	-0.004	0.302	<-0.001	0.001
	Half LOQ	0.039	0.328	-0.006	0.009
	Omit	-0.337	0.494	0.039	0.042

4.2 Tolerance Limit

To evaluate the left censored data regression tolerance limit, the limit's coverage in the simulation will be discussed first. Based on the simulation described in section 4.1, the left censored data regression model is $Y(t) = 6.7 - 0.2t + \ \varepsilon,\ \varepsilon \sim N(0,1)$. In this section, we focus on four tolerance limit settings. They are 95% tolerance limit with 95% or 99% confidence levels and 99% tolerance limit with 95% or 99% confidence levels. Then coverage probabilities of these tolerance limits are compared based on left censored data regression. The simulation consists of the following steps:

1. Set four time points, t = 6, 12, 18 and 24. Based on the left censored data regression model, generate 10 corresponding responses at each time point with $\varepsilon \sim N(0,1)$.
2. Fit the left censored data regression with the data from step 1 and LOQ=1.
3. Calculate the tolerance limits using two settings: 95% tolerance limit with 95% confidence level and the 99% tolerance limit with a 95% confidence level
4. Calculate the coverage of the tolerance limit at a given time. It is the probability of finding a response value under the tolerance limit. Since the LOQ data is the most likely to occur at the last time point than others, with time t = 24. It means that $y|t = 24 \sim N\left(\mu_{y|t=24}, 1\right),\ \mu_{y|t=24} = 1.9$
5. Repeat steps 1 thru 4 1000 times to estimate how often a certain coverage occurs.

Figure 4.1 and Figure 4.2 show the histogram plots of coverage probability for this simulation. Table 4.2 shows summary results of the coverage probability from the four tolerance limits. After running this simulation, the average coverage probability of the 95% tolerance limit with a 95% confidence level is 83%. The average coverage probability of the 99% tolerance limit

with a 95% confidence level is 92%. 99% tolerance limit with a 95% confidence level offers the highest coverage probability with 100% median coverage probability. However, in practice, the 99% tolerance limit with 99% confidence level may be too wide. It may result in too conservative withdrawal period. In order to study this, we set up the second simulation.

Table 4.2: Summary of Coverage Probability

Settings		**Coverage Probability**			
Tolerance Limit	**Confidence Level**	**25 Percentile**	**Mean**	**Median**	**75 Percentile**
95%	**95%**	0.782	0.834	0.921	1.000
	99%	0.795	0.845	0.935	1.000
99%	**95%**	0.916	0.923	1.000	1.000
	99%	0.920	0.931	1.000	1.000

We set the maximum residue level MRL=2. Based on the left censored data regression model $Y(t) = 6.7 - 0.2t + \varepsilon,\ \varepsilon \sim N(0,1)$, we can estimate the 95-percentile limit band. The time point of the intersection of MRL and the limit band is viewed as true withdrawal period. So, the true withdrawal period is 25.1 (time units). The simulation consists of the following steps:

1. Set four time points, t = 6, 12, 18 and 24. Based on the left censored data regression model, generate 10 corresponding responses at each time point with $\varepsilon \sim N(0,1)$.
2. Fit the left censored data regression with the data we got from step 1 and LOQ=1.
3. Calculate the tolerance limits using two settings: 95% tolerance limit with 95% confidence level and 99% tolerance limit with 95% confidence level

4. Calculate the withdrawal period with MRL=2.
5. Repeat steps 1 to 4 1000 times.
6. Calculate percentage of withdrawal period below and above true withdrawal period.

The simulation results are showed in Table 4.3. It is clear to see that the estimated withdrawal period increases when the tolerance limit and confidence level increase. The 95% tolerance limit with a 95% confidence level establishes the shortest withdrawal period which is 25.102. This withdrawal period is very close to the true withdrawal period. The 99% tolerance limit with a 99% confidence level establishes the longest withdrawal period which is 26.128. Depending on the research purposes, researchers can choose different tolerance limit settings in order to have moderate or conservative drug withdrawal periods.

Table 4.3: Withdrawal Period Simulation Summary

Settings		**Estimated Withdrawal Period**	
Tolerance Limit	**Confidence Level**	**Mean**	**Standard Deviation**
95%	**95%**	25.102	0.881
	99%	25.172	0.896
99%	**95%**	26.103	1.040
	99%	26.128	1.067

Assumed True Withdrawal Period is 25.

Chapter 5 - Application

In this chapter, I applied the left censored data regression and GPQ tolerance limit to establish the withdrawal period based on a publicly available data from EMA-CVMP (2016).

This data was constructed from an empirical residue depletion study on cattle treated with a veterinary drug. It was used to demonstrate the applicability of the statistical model for the estimation of withdrawal periods. The residue data are for the drug residue in the target tissue liver.

An average daily intake (ADI) of 35 μg per day for a 60 kg person has been assumed for the total residue. Then, the maximum residue level (MRLs) for the marker residue have been set at 30 μg/kg for the liver respectively（EMA-CVMP 2016). There are 48 non-missing observations. 12 liver samples are measured at time 7, 14, 21, 28 days. All liver drug residue concentration data were transformed to log scale. Then we use log (2) = 0.7 as LOQ in log scale. There are 5 observations, 10% of the dataset, below LOQ.

A left censored data regression model is fitted for this data set. The model is

$$Y(t) = 5.64 - 0.16t,$$

where $\beta_0 = 5.64, \beta_1 = -0.16$ and $\sigma = 0.95$. The 95% GPQ tolerance limit with 95% confidence interval is established. Based on MRLs = 30 μg/kg (3.4 in log scale) the withdrawal period is 25.4 days (Fig. 5.1). Based on the FDA's policy, we need to round withdrawal period to the next day (FDA-CVM 2016). So, the final withdrawal period is 26 days, meaning the drug

treated cattle should be safe for human to consume after 26 days. From Table 5.1 it can be seen that the drug residue concentrations drop below MRL (30 μg/kg) after 26 days.

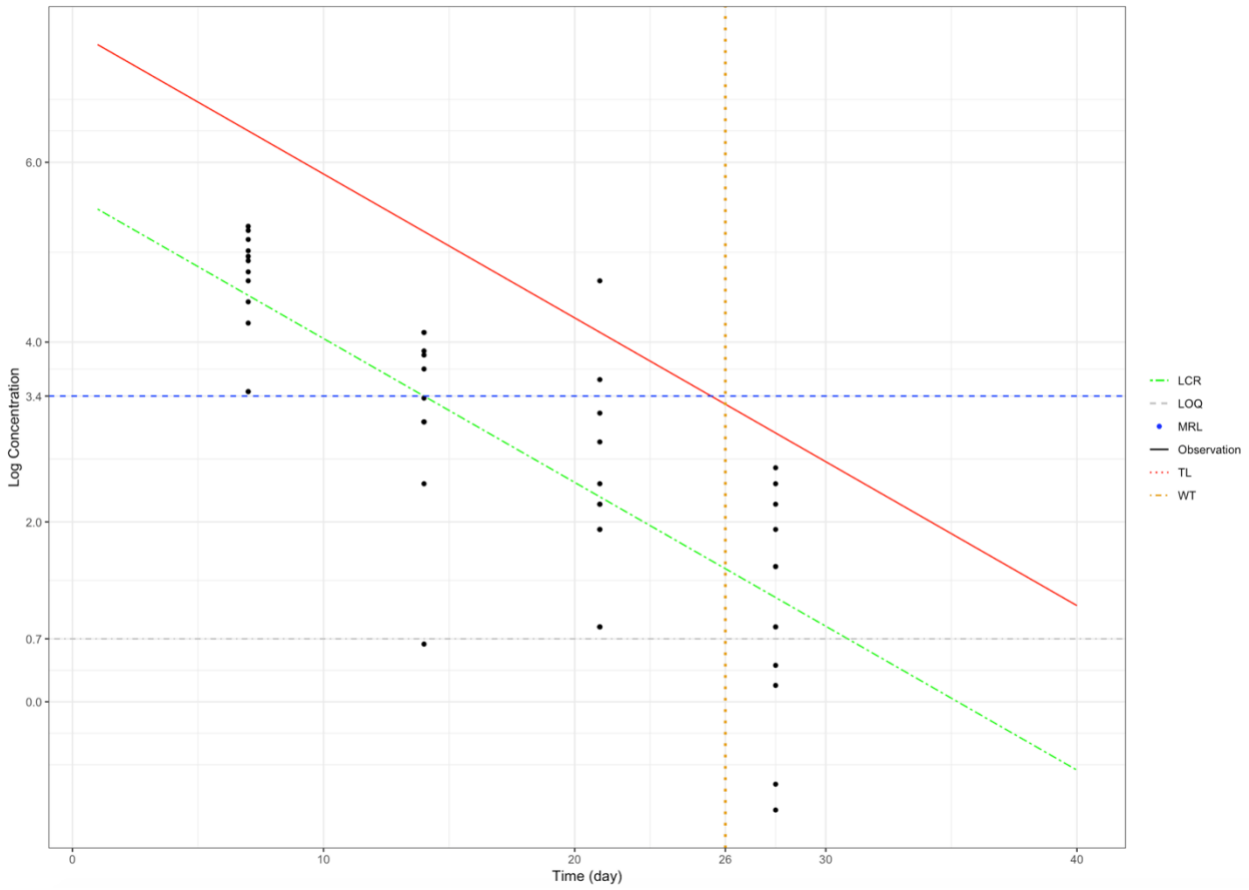

Figure 5.1: Withdrawal Period on Application Data Set. LCR is left censored data regression line. LOQ is log scaled limit of quantitation, which is 0.7. MRL is maximum residue level, which is 3.4 (log 30 =3.4). TL is 95% tolerance limit with 95% confidence interval based on left censored data regression. WT is withdrawal time, which 26 days.

Table 5.1: Drug Residue Concentrations at 24, 25, 26 and 27 Days.

	Time (Days)			
	24	25	26	27
Concentrations (ug/kg)	37.7	32.1	27.4*	23.3*

MRL is 30 ug/kg.

* indicates the drug residue below MRL.

At 21 days, there is one data point far above main group and above the upper tolerance limit. Based on the data available, we don't have enough evidence to show it is an outlier; thus, it was kept in the model as regular observation point. In practice, if we can confidently determine a certain data point is an outlier, we can exclude it. Otherwise, such data points remain in the model. However, extreme data points will have effects on the regression model. The EMA suggests fitting models with and without outliers or high leverage points and comparing the results. Perhaps in the future, a left censored data regression method can be developed to deal with high leverage observations.

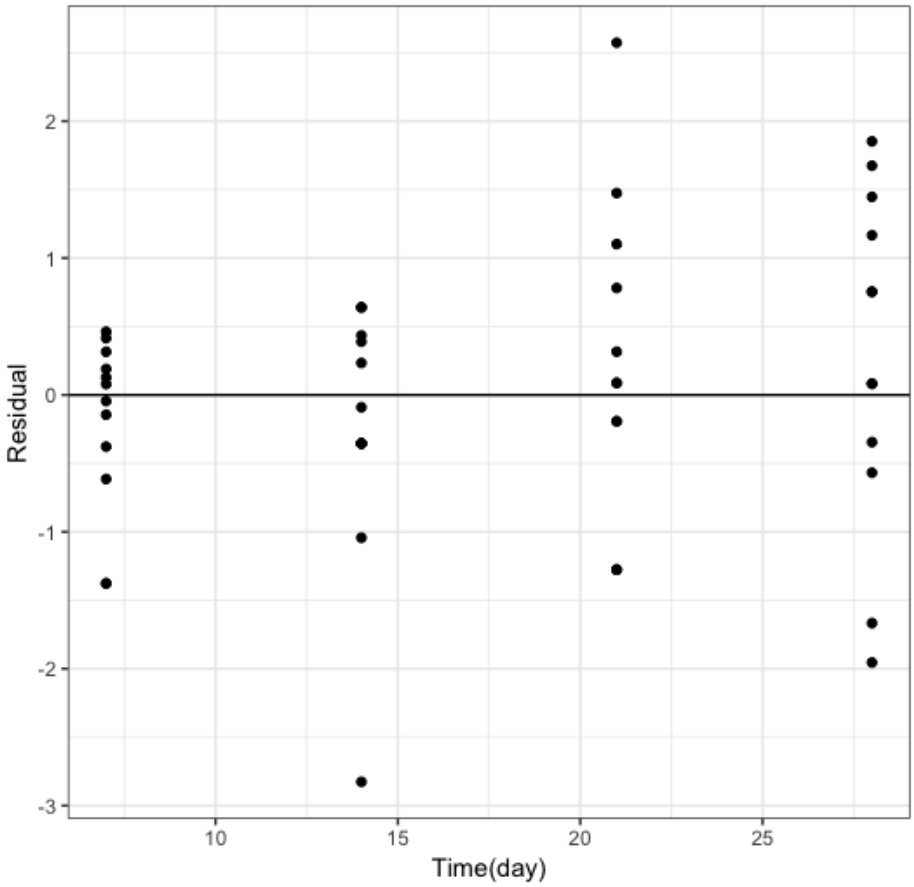

Figure 5.2: Residual Analysis. There are 12 residuals for each time point 7, 14, 21, 28 days.

In addition, Figure 5.2 shows the residual analysis plot. The residuals may not be homogenous as they appear to increase over time. The FDA and EMA do not have any guidance with respect to heterogeneous variance issues. Hence, future work might focus on a heterogenous setting for left censored data regression.

Chapter 6 - Conclusion

This dissertation study focuses on solving a real-world problem. In veterinary drug residue research, the situation where a dataset includes observations below LOQ occurs frequently. Currently, FDA recommends excluding all observations below LOQ. Considering limited research funding and resources, the loss of information resulting from this approach makes an inferior option. Based on our simulation results in section 4.1, the approach of omitting observations below LOQ resulted in high bias and RMSE of the parameter estimates. The half LOQ approach is recommended by the EMA to handle data below the LOQ. According to our simulation results, this approach performs better than the "omit" approach, which is in agreement with the findings of EMA studies (EMV-CVMP 2016). Of the three approaches considered, the proposed left censored data regression method performs the best. It has much smaller bias and RMSE of β_0 and β_1 than the half LOQ and omitting LOQ approaches, especially when a large proportion of the data falls below the LOQ.

The tolerance limit estimates of the GPQ approach have very good performance for left censored data regression. The median of coverage probability of the 95% tolerance limit with a 95% confidence level is 92%. The median of coverage probability of the 99% tolerance limit with a 95% confidence level is 100%. These results demonstrate that the GPQ approach tolerance limit can cover the population very well. In the tolerance limit simulation study, it appears that the 99% tolerance limit with a 95% confidence level may be too wide. The withdrawal period constructed by this limit is more conservative than the one established by the 95% tolerance limit with a 95% confidence level. Also, the estimated withdrawal period with 95% tolerance limit with a 95% confidence level is very close to true withdrawal period.

Also note that the maximum likelihood was used to estimate all parameters. However, MLEs of variances tend to be biased. Future work might focus on developing robust unbiased estimation approaches. In practice, it is possible that the drug residue concentration does not follow a log normal distribution over time. Future study focusing on other distributional assumptions should be undertaken.

Finally, the physiologically-based pharmacokinetic (PBPK) model has been used in veterinary drug depletion studies. Pharmacokinetic (PK) originated from the scientific basis of modern pharmacotherapy. Pharmacokinetics is the study of the rate and extent of drug transport in the body to various tissues, beginning from the time of its administration to its absorption, distribution, metabolism, and excretion, commonly abbreviated by ADME (Peters, 2012). The PBPK model was developed to capture key physiological, biochemical, and physicochemical determinants for the time course of ADME of chemicals and their metabolites in the body using a set of mathematical equations. Hence, PBPK models could be used to simulate data sets which are closer to real data than those simulated here. The PBPK model is introduced in more detail in Appendix A. Specific PBPK models are shown for complicated drug depletion processes in eggs with and without metabolism. In the future, it may be possible to develop a left censored data regression model based on possibly more reliable simulated data sets from PBPK models, especially for complex drug depletion processes.

References

Aitkin, M. (1981). A note on the regression analysis of censored data. Technometrics, 23(2), 161-163.

Améndola, C., Drton, M., & Sturmfels, B. (2015, November). Maximum likelihood estimates for Gaussian mixtures are transcendental. In International Conference on Mathematical Aspects of Computer and Information Sciences (pp. 579-590). Springer, Cham.

Ames, B. N. (1983). Dietary carcinogens and anticarcinogens. Science 221 (4617), 1256–1264.

Armbruster, D. A., & Pry, T. (2008). Limit of blank, limit of detection and limit of quantitation. The Clinical Biochemist Reviews, 29(Suppl 1), S49.

Armbruster, D. A., Tillman, M. D., & Hubbs, L. M. (1994). Limit of detection (LQD)/limit of quantitation (LOQ): comparison of the empirical and the statistical methods exemplified with GC-MS assays of abused drugs. Clinical chemistry, 40(7), 1233-1238.

Bańbura, M., & Modugno, M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics, 29(1), 133-160.

Beal, S. L. (2001). Ways to fit a PK model with some data below the quantification limit. Journal of pharmacokinetics and pharmacodynamics, 28(5), 481-504.

Beyene, T. (2016). Veterinary drug residues in food-animal products: Its risk factors and potential effects on public health. Journal of Veterinary Science & Technology 7 (1), 1–7.

Bhattacharyya, G. K. (1985). The asymptotics of maximum likelihood and related estimators based on type II censored data. Journal of the American Statistical Association, 80(390), 398-404.

Boeckel, J. N., Palapies, L., Zeller, T., Reis, S. M., von Jeinsen, B., Tzikas, S., ... & Zeiher, A. M. (2015). Estimation of values below the limit of detection of a contemporary sensitive troponin I assay improves diagnosis of acute myocardial infarction. Clinical chemistry, clinchem-2015.

Borror, C. M., Montgomery, D. C., & Runger, G. C. (1997). Confidence intervals for variance components from gauge capability studies. Quality and Reliability Engineering International, 13(6), 361-369.

Burdick, R., & Graybill, F. (1992). Confidence intervals on variance components. Marcel Decker. Inc., New York.

Clinical and Laboratory Standard Institution (2004). Protocols for Determination of Limits of Detection and Limits of Quantitation. CLSI document EP17.

Cox, D. R. (2018). Analysis of survival data. Routledge.

De Gryze, S., I. Langhans, and M. Vandebroek (2007). Using the correct intervals for prediction: A tutorial on tolerance intervals for ordinary least-squares regression. Chemometrics and intelligent laboratory systems 87 (2), 147–154.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.

Efron, B. (2018). Curvature and inference for maximum likelihood estimates. The Annals of Statistics, 46(4), 1664-1692.

EMA-CVMP (2016). Guideline on approach towards harmonization of withdrawal periods. European Medicines Agency (EMA)Committee for Medicinal Products for Veterinary Use (CVMP).

EMA/CVMP/VICH/463199/2009 (2009). VICH topic GL48: Studies to evaluate the metabolism and residue kinetics of veterinary drugs in food producing animals: marker residue depletion studies to establish product withdrawal periods. European Medicines Agency (EMA) Committee for Medicinal Products for Veterinary Use (CVMP).

FDA (1994). General Principles for Evaluating the Safety of Compounds Used in Food-Producing Animals. U.S. Department of Health and Human Services Food and Drug Administration Center for Veterinary Medicine.

FDA (2005). General Principles for Evaluating the Safety of Compounds Used in Food-Producing Animals. U.S. Department of Health and Human Services Food and Drug Administration Center for Veterinary Medicine.

FDA-CVM (2016). General Principles for Evaluating the Human Food Safety of New Animal Drugs Used in Food-Producing Animals. U.S. Department of Health and Human Services Food and Drug Administration Center for Veterinary Medicine.

Glasziou, P. P., Simes, R. J., & Gelber, R. D. (1990). Quality adjusted survival analysis. Statistics in medicine, 9(11), 1259-1276.

Graybill, F. A., & Wang, C. M. (1980). Confidence intervals on nonnegative linear combinations of variances. Journal of the American Statistical Association, 75(372), 869-873.

Hahn, G. J., & Meeker, W. Q. (2011). Statistical intervals: a guide for practitioners (Vol. 92). John Wiley & Sons.

Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69(4), 315.

Helsel, D. R. (2011). Statistics for censored environmental data using Minitab and R (Vol. 77). John Wiley & Sons.

Heskes, T. (1997). Practical confidence and prediction intervals. In Advances in neural information processing systems (pp. 176-182).

Ichihara, M., Yamamoto, A., Kakutani, N., Sudo, M., & Takakura, K. I. (2017). A Bayesian approach for estimating hexabromocyclododecane (HBCD) diastereomer compositions in water using data below limit of quantification. Environmental Science and Pollution Research, 24(3), 2667-2674.

Kalbfleisch, J. D., & Prentice, R. L. The Statistical Analysis of Failure Time Data 1980 New York. NY John Wiley & Sons.

Kelly, P. J., & Lim, L. L. Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. Statistics in medicine, 19(1), 13-33.

Krishnamoorthy, K., & Mathew, T. (2009). Statistical tolerance regions: theory, applications, and computation (Vol. 744). John Wiley & Sons.

Kurt Stange. (1971). Angewandte Statistik (Vol. II, pp. 141-143). Springer Verlag, Berlin, Heidelberg, New York.

Lawson, G. M. (1994). Defining limit of detection and limit of quantitation as applied to drug of abuse testing: striving for a consensus. Clinical chemistry, 40(7), 1218-1219.

Miller Jr, R. G. (2011). Survival analysis (Vol. 66). John Wiley & Sons.

Myers, R. H. R. H. (1990). Classical and modern regression with applications. Technical report.

Peters, S. A. (2012). Physiologically-based pharmacokinetic (PBPK) modeling and simulations: principles, methods, and applications in the pharmaceutical industry. John Wiley & Sons.

Pobiner, B. (2013). Evidence for meat-eating by early humans. Nature Education Knowledge, 4(6), 1.

Rasmussen, P. W., Staggs, M. D., Beard Jr, T. D., & Newman, S. P. (1998). Bias and confidence interval coverage of creel survey estimators evaluated by simulation. Transactions of the American Fisheries Society, 127(3), 469-480.

Schmee, J., Gladstein, D., & Nelson, W. (1985). Confidence limits for parameters of a normal distribution from singly censored samples, using maximum likelihood. Technometrics, 27(2), 119-128.

Şengül, Ü. (2016). Comparing determination methods of detection and quantification limits for aflatoxin analysis in hazelnut. Journal of food and drug analysis, 24(1), 56-62.

Senn, S., Holford, N., & Hockey, H. (2012). The ghosts of departed quantities: approaches to dealing with observations below the limit of quantitation. Statistics in medicine, 31(30), 4280-4295.

Thompson, M. L., & Nelson, K. P. (2003). Linear regression with Type I interval-and left-censored response data. Environmental and Ecological Statistics, 10(2), 221-230.

Tsui, K. W., & Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. Journal of the American Statistical Association, 84(406), 602-607

United States Census Bureau (USCB) (2014, December). Census bureau projects u.s. and world populations on new year day. https://www.census.gov/newsroom/pressreleases/2014/cb14-tps90.html. Release Number: CB14-TPS.90.

United States Department of Agriculture (USDA) Foreign Agricultural Service (2017, April). Livestock and poultry: World markets and trade. http://usda.mannlib.cornell.edu/usda/current/livestock-poultryma/livestock-poultry-ma-04-11-2017.pdf.

USEPA (U.S. Environmental Protection Agency). 2009. Integrated Science Assessment for Particulate Matter (Final Report). EPA/600/R08/139F. Washington, DC:U.S. EPA.

USDA Economic Research Service (2017a). Cattle and beef statistics and information. https://www.ers.usda.gov/topics/animal-products/cattle-beef/statisticsinformation.aspx.

USDA Economic Research Service (2017b, April). Poultry and eggs. https://www.ers.usda.gov/topics/animal-products/poultry-eggs/.

Weerahandi, S. (1995). Generalized confidence intervals. In Exact Statistical Methods for Data Analysis (pp. 143-168). Springer, New York, NY.

Wingo, D. R. (1993). Maximum likelihood methods for fitting the Burr type XII distribution to multiply (progressively) censored life test data. Metrika, 40(1), 203-210.

Yan, D. (2014). Estimation of an upper tolerance limit for small-samples containing observations below the limit of quantitation. Master Report. Kansas State University.

21CFR520.905a (2017, April). Code of federal regulations (cfr), title 21, section 520.905a fenbendazole suspension.

Appendix A - Physiologically Based Pharmacokinetic (PBPK) Model of Drug Residue Depletion

Pharmacokinetic (PK) is from the scientific basis of modern pharmacotherapy (Meibohm and Derendorf, 1997). Pharmacokinetics is the study of the rate and extent of drug transport in the body to the various tissues, beginning from the time of its administration to its absorption, distribution, metabolism, and excretion (ADME) (Peters, 2012). Put simply, it is a study to describe "what the body does to the drug" (Meibohm and Derendorf, 1997). Hence, PK modeling is the quantitative modeling of the time course of ADME of chemicals and their metabolites in the body using a set of mathematical equations. In order to capture key physiological, biochemical, and physicochemical determinants in model, the physiologically-based pharmacokinetic (PBPK) model was developed. It is built using a similar mathematical framework as the PK model, but it is parameterized using known physiology and consists of a larger number of compartments, which correspond to the different organs or tissues in the body (Andersen et al., 1987). PBPK model provides more specific model prediction by species-specific physiological parameters and chemical-specific parameters than the general PK model.

A.1 Mathematical Description on Chemical Movement

Here is an example of a simple two compartments PK model structure (see Figure A.1). It describes how one certain chemical distribute in the different body compartments. Those body compartments can be theoretical compartments or physiological organs or tissues. Normally, researchers use a rate equation to describe the movement of chemical into a compartment, out of a compartment, or between compartments.

The first order (the rate is proportional to the concentration) rate of one chemical movement process is $Rate = K \times C^1 = K \times C$, which can be written as $\frac{dC}{dt} = K \times C$, where C is concentration. For example, the rate of the chemical move from compartment 1 to compartment 2 is $\frac{dC}{dt} = K_{12} \times C_1$ by the first-order process, where C_1 is the compartment 1 concentration. That means the chemical concentration in compartment 1 changes by $K_{12} \times C_1$ per time unit causing by compartment 2.

There is another method to describe the chemical movement process, which the rate is constant and does not depend on the concentration. It is called zero-order process. The rate is $\frac{dC}{dt} = K \times C^0$. The way to choose which method depends on chemical characteristics, research assumptions, etc.

If there are more than one chemical involved in the process, we can use saturable method, which can describe metabolism transport across membranes. We can use this in the drugs that has metabolic process. The saturable process rate is $\frac{dC}{dt} = \frac{V_{max}+C}{C+K_m}$, where V_{max} is the maximum rate, $K_m$is the substrate concentration that is required for the reaction to occur at half of V_{max}.

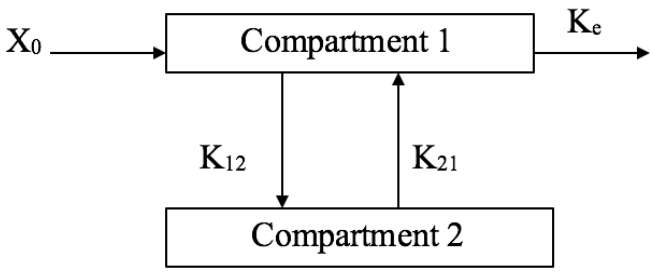

Figure A.6.1: Two compartments PK model. X0 = Dose of drug (mg), Ke = elimination rate constant (h^{-1}), K12= rate of transfer from compartment 1 to compartment 2 (h^{-1}), K21= rate of transfer from compartment 2 to compartment 1 (h^{-1}).

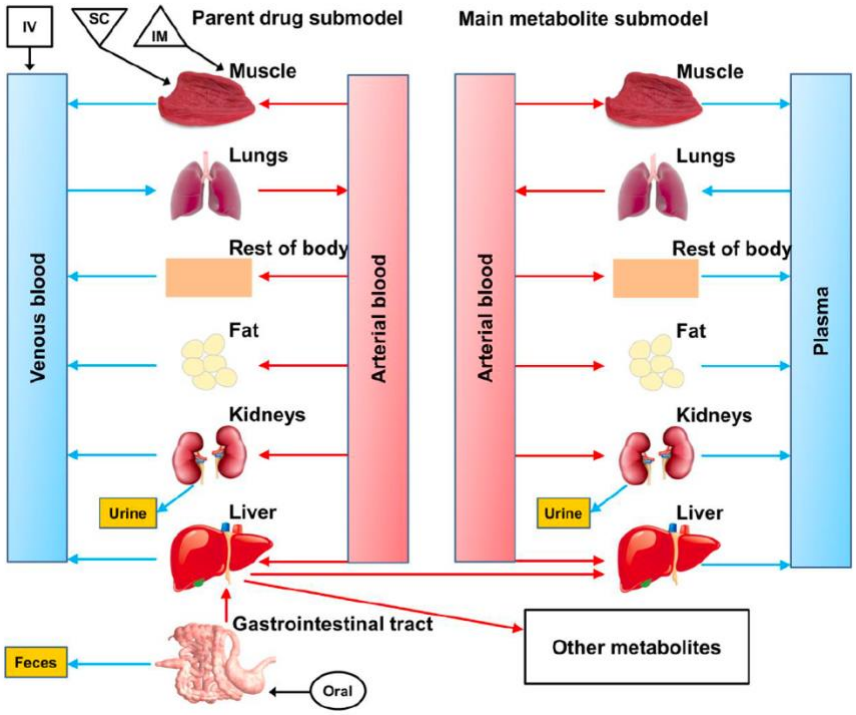

Figure A.6.2: An example of PBPK model structure. IV is intravenous, IM is intramuscular, SC is subcutaneous. (Meibohm and Derendorf, 1997)

The PBPK model structure is more complicated than PK model structure. A PBPK model structure example is showed in Figure A.2. From this figure, it is easier to understand PBPK model. It describes the absorption, distribution, metabolism, and elimination (ADME) of

environmental chemicals, drugs, or nonmaterial in an organism based on interrelationships among key physiological, biochemical, and physicochemical determinants using mathematical equations. The compartments are blood, liver, kidneys, muscle, lungs, etc.. We can include species-specific physiological parameters (e.g., body weight, cardiac output, organ mass, etc.) and chemical-specific parameters (e.g., partition coefficients, permeability coefficients, metabolic rate constants, etc.) in the model to increase model prediction accuracy. Overall, no matter which model to use, the chemical transport rates are all based on the three calculate methods we introduced above in our research.

A.2 Eggs Specific PBPK Model

Reproduction is a complex process. Chickens lay an egg roughly every 24 hours (Goetting et al., 2011). It takes several days for each egg to develop in vivo (Etches et al., 1996). Eggs consist of three main compartments: yolk, albumen, and shell. Of these three parts, the egg yolk has the longest development time. The yolk and albumen develop at different stages of the egg formation process. The formation of the yolk takes 8 to 10 days. However, the formation of the albumen only takes approximately 10 hours (Hekman and Schefferlie, 2011) following after yolk formation process. An active laying hen can have several yolk follicles at different development stages before ovulation at the same time. After ovulation, the albumen proteins are deposited in the magnum and electrolytes. Water are "pumped" in to the albumen in the distal parts of the oviduct (isthmus, uterus) (Schefferlie and Hekma,n 2016). The egg shell is added after this process (Etches et al., 1996). The egg development process is similar across species (Whittow, 1999).

The detailed diagrams of a chicken reproduction system and a chicken egg is shown in Figure A.3 and Figure A.4. Drug deposits in the egg yolk rapidly accumulate during the rapid growth stage in 8 to 10 days. After egg yolk gets developed well, egg albumen accumulates drug residue in the oviduct for roughly 10 hours. Since humans normally do not consume egg shell, egg yolk and egg albumen are the two residue accumulation sites of our research concerns. The egg yolk and egg albumen are developed in a different time range (egg albumen starts to develop after egg yolk form process done), so these two egg components are distinct in physiological process. In addition, it has been reported that drug residue profiles are different between egg yolk

and egg albumen (Chu et al., 2000). Therefore, the drug disposition into each component (yolk and albumen) should be described separately in models.

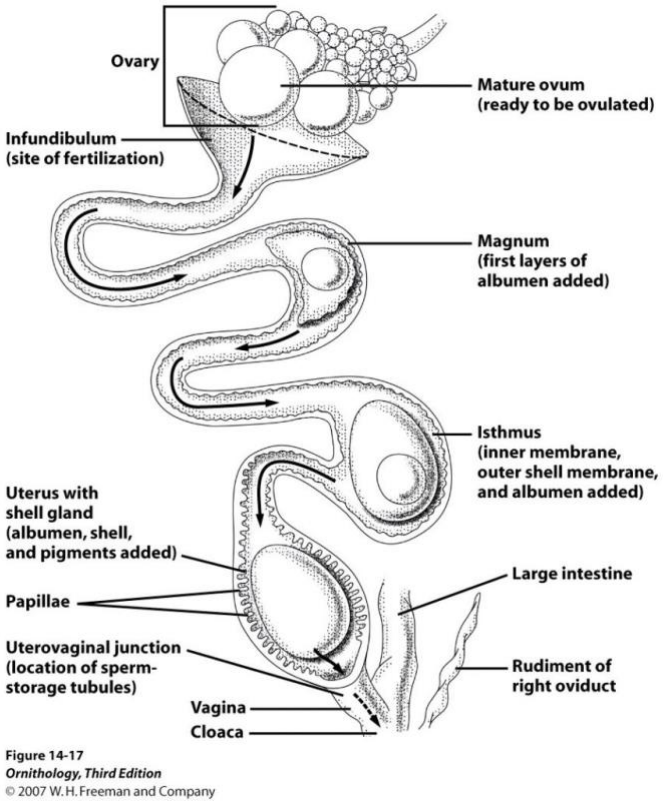

Figure A.6.3: A typical bird reproduction system. Source: http://www.bhwt.org.uk/produce/

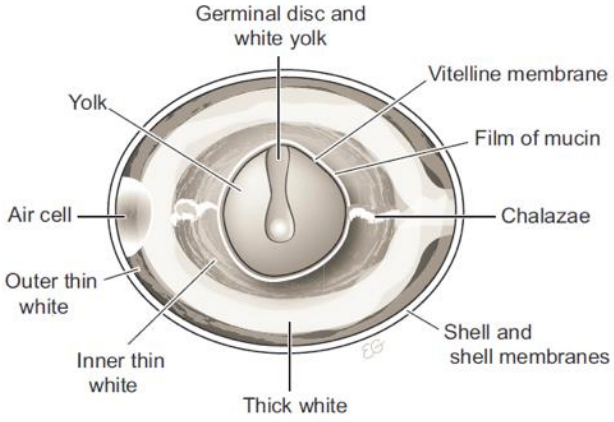

Figure A.6.4: A detailed illustration of developing egg components (Goetting et al., 2011)

Actually, it has been proven that the traditional pharmacokinetic (PK) model, which is used to describe the time dependent course of drug residue in plasma and tissues, fails to in egg drug residue depletion studies (Schefferlie and Hekman, 2016). It is clear that drug residue concentration in eggs is highly related to egg formation stage, which is described in the last paragraph. The model should correspond to this factor. The disposition of drugs into the egg yolk and egg albumen is through a filtration-process rather than through a concentration-gradient driven process (Hekman and Schefferlie, 2011). This indicates the drug movement does not depend on concentration difference. It continues moving to egg if there is still any drug in plasma. Hence, the physiology of egg yolk and albumen formation factor must be involved in the model.

Moreover, plasma concentration is recognized as an important influence on drug disposition by many researchers (Hekman and Schefferlie, 2011). Therefore, drug residue models of eggs should consider the yolk and albumen formation stage, plasma concentration and physicochemical properties of the drug. In 2011, P.Hekman and G.J.Schefferlie developed a PBPK model to meet all these requirements. The model structure is shown in Figure A.5.

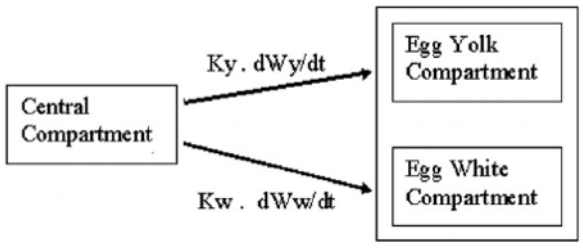

Figure A.6.5: PBPK model structure for drug residue in egg (Hekman and Schefferlie, 2011)

And, the transport rate equation of drug into egg yolk and egg white is:

$$\frac{dQ_y}{dt} = C_P(t) \times K_y \times \frac{dW_y}{dt} \qquad (A.1)$$

$$\frac{dQ_w}{dt} = C_P(t) \times K_w \times \frac{dW_w}{dt} \qquad (A.2)$$

where, $\frac{dQ_y}{dt}$ and $\frac{dQ_w}{dt}$ are the rate of drug deposition into yolk and albumen. C_P is plasma concentration. Ky and Kw are transport constants into yolk and albumen (white). $\frac{dW_y}{dt}$ and $\frac{dW_w}{dt}$ are the yolk and albumen (white) formation rate (g/day), describing its growth stage.

We can develop PBPK model for drug residue depletion study in eggs based on this.

In order to estimate$\frac{dQ_y}{dt}$ and $\frac{dQ_w}{dt}$, we need to understand each part ($C_P(t)$, Ky, Kw, $\frac{dW_y}{dt}$ and $\frac{dW_w}{dt}$) and know how to calculate them. $C_P(t)$ is a plasma concentration function depending on time, which C_P is experiment measured observations. We don't have to worry about this part. Next are the egg formation rates. Let's talk about yolk formation ($\frac{dW_y}{dt}$) rate first.

Based on published papers (Geertsma et al. 1987; Kan and Petz 2000), the rate of egg yolk formation ($\frac{dW_y}{dt}$) is not linear depending on time. The rapid yolk growth period is an exponential distribution during the last 8 - 10 days before ovulation (Geertsma et al., 1987). Hekman and Schefferlie (2011) claimed that the time dependent of egg yolk weight was sigmoidal shape (shown in Figure A.6). At time t = tlay (an egg laid time), the yolk weight can be written as:

$$W_y(t) = \frac{A_y}{1 + e^{-s\times(t-(tlay-tsig-tlag))}} \qquad (A.3)$$

where, tlay is the time of egg lay, tlag is time between ovulation and egg lay (approximately one day), Wy is weight of yolk at given time, Ay is apparent maximum yolk weight, s is maximum daily yolk growth rate constant and tsig is pre-ovulation time of maximum yolk growth rate (approximately 1-3 days before ovulation). Ay, s and tsig can be estimated by experiment data from Geertsema et al. 1987. In this case, it has Ay = 25, s = 1 and tsig = 2.

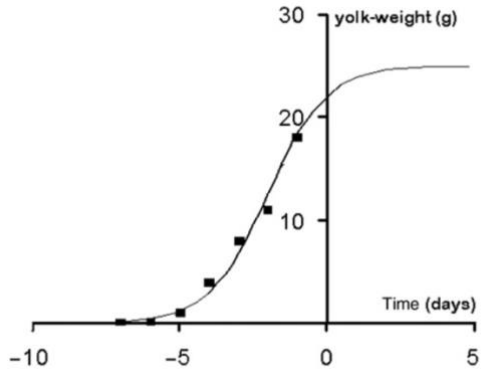

Figure A.6.6: Egg yolk formation rate by Hekman and Schefferlie (2011) using data from Geertsema et al. 1987.Time 0 is ovulation time. Before time 0, the yolk is rapidly growing. The weight of egg yolk stops increasing shortly after ovulation.

After taking first derivative of sigmoid function Wy(t), the egg yolk formation rate can be written as:

$$\frac{dW_y}{dt} = \frac{A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{(1+e^{-s\times(t-(tlay-tsig-tlag))})^2} \qquad (A.4)$$

Hence, inserting function A.4 in to function A.1, the rate of drug deposition into yolk is:

$$\frac{dQ_y(t)}{dt} = \frac{C_P(t) \times K_y \times A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{(1+e^{-s\times(t-(tlay-tsig-tlag))})^2} \qquad (A.5)$$

Since the drug only goes into yolk without coming out (Figure A.5 and the yolk stops development after ovulation, the accumulated amount of drug in yolk when it laid should be integral of $\frac{dQ_y(t)}{dt}$ with time interval [0, tlay − tlag], which can be written as:

$$Q_y(tlay) = \int_{t=0}^{t=tlay-tlag} \frac{dQ_y(t)}{dt} dt$$

$$= K_y \times A_y \times s \times \int_{t=0}^{t=tlay-tlag} \frac{C_P(t) \times K_y \times A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{(1+e^{-s\times(t-(tlay-tsig-tlag))})^2} dt \qquad (A.6)$$

Also, based on function A.3, the yolk weight at egg laying time is (when t = tlay):

$$W_y(tlay) = \frac{A_y}{1+e^{-s\times(t-(tlay-tsig-tlag))}} = \frac{A_y}{1+e^{-s\times(tsig+tlag))}} \qquad (A.7)$$

Then, the egg yolk drug concentration at laying time can be calculated with $\frac{Q_y(tlay)}{W_y(tlay)}$, which can be represented by:

$$C_y(tlay) = \frac{Q_y(tlay)}{W_y(tlay)}$$

$$= \frac{K_y \times A_y \times s \times \int_{t=0}^{t=tlay-tlag} \frac{C_P(t) \times K_y \times A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{\left(1+e^{-s\times(t-(tlay-tsig-tlag))}\right)^2} dt}{\frac{A_y}{1+e^{-s\times(tsig+tlag))}}}$$

$$= K_y \times s \times \left(1+e^{-s\times(tsig+tlag))}\right)$$

$$\times \frac{\int_{t=0}^{t=tlay-tlag} C_P(t) \times e^{-s\times(t-(tlay-tsig-tlag))}}{\left(1+e^{-s\times(t-(tlay-tsig-tlag))}\right)^2} dt \qquad (A.8)$$

where, Cy(tlay) is observed data from experiment. Hence, Ky can be estimated by function A.8.

Even though the formation rate of egg albumen is different than yolk, the procedure to get formation rate is quite similar. The albumen is only deposited in egg after ovulation. The only difference is that albumen only takes several hours to rapid formation and does not exist

before ovulation. Hence, when we take the derivative of the function, the time interval should start from after ovulation, which should be bigger than 8 days. Similar as Ky estimation process, Kw can be estimated by flowing steps.

We know egg albumen only takes 8-10 hours to be added onto egg after ovulation. This process is much shorter than egg yolk. It is not necessary to use sigmoid function here. We can just use a constant rate to describe albumen formation rate, which can be written as:

$$\frac{dW_w}{dt} = A_w \qquad (A.9)$$

where Aw is egg albumen constant formation rate (g/day). Then, function A.2 can be developed to:

$$\frac{dQ_w}{dt} = C_P(t) \times K_w \times A_w \qquad (A.10)$$

As same as drug moves into yolk, the drug does not come out if it goes into albumen. Also, the albumen only develops after ovulation. So, the accumulated amount of drug in albumen when it laid should be integral of $\frac{dQ_w(t)}{dt}$ with time interval [tlay − tlag, tlay − tlag + talbumen], which can be written as:

$$Q_w(tlay) = \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} \frac{dQ_w(t)}{dt} dt$$

$$= K_w \times A_w \times \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} C_P(t)dt \qquad (A.11)$$

where talbumen is albumen development time (8-10 hour). Since the egg albumen formation rate is constant Aw (g/day), the albumen weight at laying time is

$$W_w(tlay) = A_w \times (tlay - tlag + talbumen) - A_w \times (tlay - tlag)$$

$$= A_w \times talbumen \qquad (A.12)$$

Then, the drug concentration in egg albumen at laying time should be $\frac{Q_w(tlay)}{W_w(tlay)}$, which can be written as:

$$C_w(tlay) = \frac{Q_w(tlay)}{W_w(tlay)}$$

$$= \frac{K_w \times A_w \times \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} C_P(t)dt}{A_w \times talbumen}$$

$$= \frac{K_w}{talbumen} \times \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} C_P(t)dt \qquad (A.13)$$

Hence, Kw can be estimated by function A.13 based on observed plasma and drug concentration in albumen data.

A.3 PBPK Model for Non-metabolic Drug Residue Depletion in Eggs

Actually, for on metabolic drug, the PBPK model structure is shown in Figure A.7.

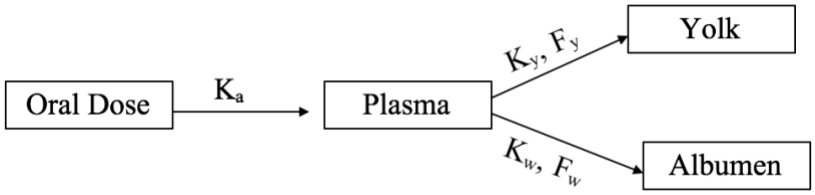

Figure A.6.7: Physiologically-based pharmacokinetic (PBPK) model structure of simple (non-metabolic) drugs in eggs. Ka is absorption rate. Ky and Kw are transportation constants from plasma into yolk and albumen (white). Fy and Fw are formation rate of yolk and albumen (white), where $F_y = \frac{dW_y}{dt}$ and $F_w = \frac{dW_w}{dt}$. Kr is transportation constants into rest of body compartments from plasma. Kp is transportation constants into plasma from rest of body compartments.

Based on the discussion in last section and in Hekman and Schefferlie (2011), we know the rate of drug deposition into egg yolk and albumen are:

$$\frac{dQ_y}{dt} = C_P(t) \times K_y \times F_y \qquad (A.14)$$

$$\frac{dQ_w}{dt} = C_P(t) \times K_w \times F_w \qquad (A.15)$$

where $F_y = \frac{dW_y}{dt}$ and $F_w = \frac{dW_w}{dt}$. Then, based on function (A.4) and (A.5), we can have

$$F_y = \frac{dW_y}{dt} = \frac{A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{(1+e^{-s\times(t-(tlay-tsig-tlag))})^2}$$

$$F_w = \frac{dW_w}{dt} = A_w$$

where tlay is the time of egg lay, tlag is time between ovulation and egg lay (approximately one day), Wy is weight of yolk at given time, Ay is apparent maximum yolk weight, s is maximum daily yolk growth rate constant and tsig is pre-ovulation time of maximum yolk growth rate (approximately 1-3 days before ovulation). Ay, s and tsig can be estimated as Ay = 25, s = 1 and tsig = 2 by experiment data from Geertsema et al., (1987). Aw is egg albumen constant formation rate (g/day). Also, based on function (A.8) and (A.13) we already know how to estimate parameter Ky and Kw by

$$C_y(tlay) = K_y \times s \times \left(1 + e^{-s\times(tsig+tlag))}\right) \frac{\int_{t=0}^{t=tlay-tlag} C_P(t) \times e^{-s\times\left(t-(tlay-tsig-tlag)\right)}}{\left(1 + e^{-s\times\left(t-(tlay-tsig-tlag)\right)}\right)^2} dt$$

$$C_w(tlay) = \frac{K_w}{talbumen} \times \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} C_P(t)dt$$

Moreover, Ka, Kr and Kp can be estimate by the model. ke= 0.28/hour with SD ± 0.09 (Souza et al., 2017). The only thing we need is the data right now. The experiment data we have does not have plasma drug concentration. However, we find a set of useable data with plasma, egg yolk and egg albumen concentration with oral administration of a single dose of the certain veterinary drug (1 mg/kg PO once) (Souza et al., 2017). So, we can build the PBPK model with it and evaluate the model by research data. The PBPK model is built and ran by Berkley Mandan (version 9.0.127) and R (Version 1.0.153 − c 2009-2017 RStudio, Inc.).

A.4 PBPK Model for Metabolic Drug Residue Depletion in Eggs

For certain drugs that has two metabolites: M1 and M2. Hence, PBPK model for this kind of drug in laying hen eggs is more complex than non-metabolic drugs. The model structure might be more complicated (Figure A.8).

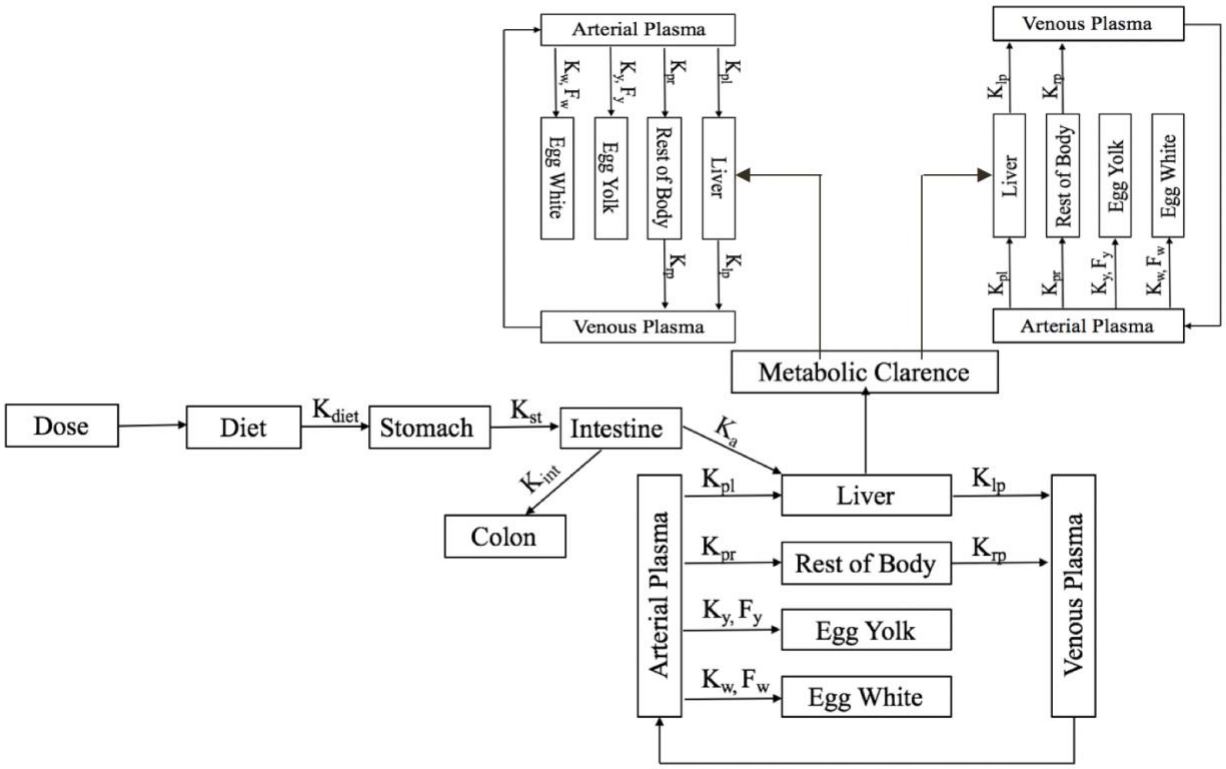

Figure A.6.8: Physiologically-based pharmacokinetic (PBPK) model structure in eggs. The bottom part is for parent drug Upper left is M1 and upper right is M2. Ky and Kw are transportation constants from plasma into yolk and albumen (white). Fy and Fw are formation rate of yolk and albumen (white), where $F_y = \frac{dW_y}{dt}$ and $F_w = \frac{dW_w}{dt}$. Kdiet is transportation constants from diet into stomach. Kst is transportation constants from stomach into intestine. Kint is elimination rate from body to colon.

As same as PBPK model of the non-metabolic drugs, we will use same estimate structure to get Fy, Fw, Ky and Kw. $F_y = \frac{dW_y}{dt} = \frac{A_y \times s \times e^{-s\times(t-(tlay-tsig-tlag))}}{(1+e^{-s\times(t-(tlay-tsig-tlag))})^2}$, $F_w = \frac{dW_w}{dt} = A_w$, $C_y(tlay) = K_y \times s \times \left(1 + e^{-s\times(tsig+tlag)}\right) \frac{\int_{t=0}^{t=tlay-tlag} C_P(t) \times e^{-s\times(t-(tlay-tsig-tlag))}}{\left(1+e^{-s\times(t-(tlay-tsig-tlag))}\right)^2} dt$ and $C_w(tlay) = \frac{K_w}{talbumen} \times \int_{t=tlay-tlag}^{t=tlay-tlag+talbumen} C_P(t)dt$. For other PBPK model rate, we can get from literature reviews.

A.5 Reference

Andersen, M., r. Clewell, HJ, M. Gargas, F. Smith, and R. Reitz (1987). Physiologically based pharmacokinetics and the risk assessment process for methylene chloride. Toxicology and applied pharmacology 87 (2), 185–205.

Chu, P.-S., D. J. Donoghue, and B. Shaikh (2000). Determination of total 14c residues of sarafloxacin in eggs of laying hens. Journal of agricultural and food chemistry 48 (12), 6409–6411.

EMEA/MRL/236 (1997, June). Committee for veterinary medicinal products meloxicam summary report (1).

Etches, R. J. et al. (1996). Reproduction in poultry. CAB international.

Geertsma, M., J. Nouws, J. Grondel, M. Aerts, T. Vree, and C. Kan (1987). Residues of sulphadimidine and its metabolites in eggs following oral sulphadimidine medication of hens. Veterinary Quarterly 9 (1), 67–75.

Goetting, V., K. Lee, and L. Tell (2011). Pharmacokinetics of veterinary drugs in laying hens and residues in eggs: a review of the literature. Journal of veterinary pharmacology and therapeutics 34 (6), 521–556.

Kan, C. A. and M. Petz (2000). Residues of veterinary drugs in eggs and their distribution between yolk and white. Journal of Agricultural and Food Chemistry 48 (12), 6397–6403.

Hekman, P. and G. J. Schefferlie (2011). Kinetic modelling and residue depletion of drugs in eggs. British poultry science 52 (3), 376–380.

Meibohm, B. and H. Derendorf (1997). Basic concepts of pharmacokinetic/pharmacodynamic (pk/pd) modelling. International journal of clinical pharmacology and therapeutics 35 (10), 401–413.

Peters, S. A. (2012). Physiologically-based pharmacokinetic (PBPK) modeling and simulations: principles, methods, and applications in the pharmaceutical industry. John Wiley & Sons.

Schefferlie, G. and P. Hekman (2016). Prediction of the residue levels of drugs in eggs, using physicochemical properties and their influence on passive diffusion processes. Journal of veterinary pharmacology and therapeutics 39 (4), 381–387.

Souza, M. J., J. B. Bergman, M. S. White, K. I. Gordon, L. E. Gerhardt, and S. K. Cox (2017). Pharmacokinetics and egg residues after oral administration of a single dose of meloxicam in domestic chickens (gallus domesticus). American Journal of Veterinary Research 78 (8), 965–968.

Whittow, G. C. (1999). Sturkie's avian physiology. Academic press.

Appendix B - R Code

```
# Simulation: Left censored data regression
## Omit
#### This is an example for sample size =24, n=1000, 5% data below LOQ
##### For sample size 32, 40, 48, the points will be 8, 10, 12, nrow=32, 40, 28
##### For 10% or 15% data below LOQ, the LOQ=1.542 or 2
points <- 6
nrow<- 24
mean1 <- 5.3322
mean2 <- 4.1934
mean3 <- 3.0546
mean4 <- 1.9158
sd <- 1
LOQ <- 1
result <- data.frame(matrix(nrow = 1000, ncol = 2))
colnames(result) <- c("i", "beta0", "beta1")
for (sim in 1:1000) {
  df <- matrix(, nrow = nrow, ncol = 3)
  df[,1] <- rep(1,2nrow)
  df[,2] <- rep(c(6,12,18,24), each=points)
  set.seed(sim)
  data1 <-  rnorm(n=points,mean=mean1,sd=sd)
  set.seed(sim)
  data2 <-  rnorm(n=points,mean=mean2,sd=sd)
  set.seed(sim)
  data3 <-  rnorm(n=points,mean=mean3,sd=sd)
  set.seed(sim)
  data4 <-  rtruncnorm(n=points,mean=mean4,sd=sd)
  df[,3] <- c(data1,data2,data3,data4)
  df <-data.frame(df)
```

```
  df <-subset(df, X3 > LOQ)
  X <- df[,1:2]
   lm <-lm(df$X3~ df$X2)
  betas.1 <- as.matrix(coef(lm))
  result[sim, 1] <- sim
  result[sim, 2] <- betas.1[1,1]
  result[sim, 3] <- betas.1[2,1]
}
beta0here <- 6.471
beta1here <- -0.1898
beta0bias <- mean(result$X2)-beta0here
beta0RMSE <- result %>%
  dplyr::mutate(newcol = (X2 - beta0here) ^ 2 ) %>%
  dplyr::summarize(rmse = sqrt(sum(newcol) / 1000))
beta1bias <- mean(result$V3)-beta1here
beta1RMSE <- result %>%
  dplyr::mutate(newcol1 = (V3 -beta1here) ^ 2 ) %>%
  dplyr::summarize(rmse1 = sqrt(sum(newcol1) / 1000))
summarize_result5 <- data.frame(matrix(nrow = 2, ncol = 2))
colnames(summarize_result5) <- c("beta0","beta1")
rownames(summarize_result5) <- c("Bias", "RMSE")
summarize_result5[1,1] <- beta0bias
summarize_result5[1,2] <- beta1bias
summarize_result5[2,1] <- beta0RMSE
summarize_result5[2,2] <- beta1RMSE
summarize_result5


## Half LOQ
#### This is an example for sample size =24, n=1000, 5% data below LOQ
##### For sample size 32, 40, 48, the points will be 8, 10, 12
```

```
##### For 10% or 15% data below LOQ, the LOQ=1.542 or 2
points <- 6
mean1 <- 5.3322
mean2 <- 4.1934
mean3 <- 3.0546
mean4 <- 1.9158
sd <- 1
LOQ <- 1
nrow<- 24
result <- data.frame(matrix(nrow = 1000, ncol = 2))
colnames(result) <- c("i", "beta0", "beta1")
for (sim in 1:1000) {
  df <- matrix(, nrow = nrow, ncol = 3)
  df[,1] <- rep(1,nrow)
  df[,2] <- rep(c(6,12,18,24), each=points)
  set.seed(sim)
  data1 <-  rnorm(n=points,mean=mean1,sd=sd)
  set.seed(sim)
  data2 <-  rnorm(n=points,mean=mean2,sd=sd)
  set.seed(sim)
  data3 <-  rnorm(n=points,mean=mean3,sd=sd)
  set.seed(sim)
  data4 <-  rtruncnorm(n=points,mean=mean4,sd=sd)
  df[,3] <- c(data1,data2,data3,data4)
  X <- df[,1:2]
  df <-data.frame(df)
  df$Y <- ifelse (df$X3 > LOQ, df$X3, LOQ/2 )
  lm <-lm(df$Y~ df$X2)                                    # by treat censored data as
uncensored data
  betas.1 <- as.matrix(coef(lm)
  result[sim, 1] <- sim
```

```
  result[sim, 2] <- betas.1[1,1]
  result[sim, 3] <- betas.1[2,1]
}
beta0here <- 6.471
beta1here <- -0.1898
beta0bias <- mean(result$X2)-beta0here
beta0RMSE <- result %>%
  dplyr::mutate(newcol = (X2 - beta0here) ^ 2 ) %>%
beta1bias <- mean(result$V3)-beta1here
beta1RMSE <- result %>%
  dplyr::mutate(newcol1 = (V3 -beta1here) ^ 2 ) %>%
  dplyr::summarize(rmse1 = sqrt(sum(newcol1) / 1000))
summarize_result5 <- data.frame(matrix(nrow = 2, ncol = 2))
colnames(summarize_result5) <- c("beta0","beta1")
rownames(summarize_result5) <- c("Bias", "RMSE")
summarize_result5[1,1] <- beta0bias
summarize_result5[1,2] <- beta1bias
summarize_result5[2,1] <- beta0RMSE
summarize_result5[2,2] <- beta1RMSE

## LCR
#### This is an example for sample size =24, n=1000, 5% data below LOQ
##### For sample size 32, 40, 48, the points will be 8, 10, 12
##### For 10% or 15% data below LOQ, the LOQ=1.542 or 2
points <- 6
mean1 <- 5.3322
mean2 <- 4.1934
mean3 <- 3.0546
mean4 <- 1.9158
sd <- 1
LOQ <- 1
```

```
nrow<- 24
result <- data.frame(matrix(nrow = 1000, ncol = 2))
colnames(result) <- c("i", "beta0", "beta1")
for (sim in 1:1000) {
  df <- matrix(, nrow = nrow, ncol = 3)
  df[,1] <- rep(1,nrow)
  df[,2] <- rep(c(6,12,18,24), each=points)
  set.seed(sim)
  data1 <-  rnorm(n=points,mean=mean1,sd=sd)
  set.seed(sim)
  data2 <-  rnorm(n=points,mean=mean2,sd=sd)
  set.seed(sim)
  data3 <-  rnorm(n=points,mean=mean3,sd=sd)
  set.seed(sim)
  data4 <-  rtruncnorm(n=points,mean=mean4,sd=sd)
  df[,3] <- c(data1,data2,data3,data4)
  X <- df[,1:2] # X with 1 in 1st column
  lm.1<-lm(df[,3]~ df[,2])
  betas.1 <- as.matrix(coef(lm.1))
  sigma.1 <- sigma(lm.1)
  i<-1
  # step 2: While loop for E step and M step
  while(T){
    ### E Step###
    # Calculate expextation of sufficient statistics E(∑〖x_ij y_i〗) and E(∑〖 y_i^2 〗) with
consored y_i replaced
    df <-data.frame(df)
    df$mu.1 <- betas.1[1,1]+ df$X2 * betas.1[2,1]
    df$Y.1 <- ifelse (df$X3 > LOQ, df$X3,
                      df$mu.1 - sigma.1 * (dnorm(((LOQ-df$mu.1)/sigma.1), 0,1))/(pnorm(((LOQ-
df$mu.1)/sigma.1), 0, 1)))
```

```
    ### M-Step for beta ###
    lm.2<-lm(Y.1 ~ X2, data=df) # using detected y and estimated censored y
    betas.2 <- as.matrix(coef(lm.2))
    ### M-Step for sigma ###
    detect.df <- filter(df,X3 > LOQ)
    censored.df<-filter(df,X3 <= LOQ)
    sigma.square.2 <- 1/(nrow(df)) * (sum((detect.df$Y.1-detect.df$mu.1)^2)+ sigma.1^2 *
                                sum(1 - ((LOQ-censored.df$mu.1)/sigma.1)*
                                    (dnorm(((LOQ-censored.df$mu.1)/sigma.1), 0,1))/
                                    (pnorm(((LOQ-censored.df$mu.1)/sigma.1), 0, 1))))
    sigma.2 <- sqrt(sigma.square.2)
    #Stopping Rule
    if (sigma.2 - sigma.1 < 1e-6 & t(betas.2-betas.1)%*%(betas.2-betas.1) < 1e-6 | i>1000) break
    #update initial value for next iteration
    betas.1 <- betas.2
    sigma.1 <- sigma.2
    #Print information for each iteration
    i<-i+1
  }
  result[sim, 1] <- sim
  result[sim, 2] <- betas.1[1,1]
  result[sim, 3] <- betas.1[2,1]
}
beta0here <- 6.471
beta1here <- -0.1898
beta0bias <- mean(result$X2)-beta0here
beta0RMSE <- result %>%
  dplyr::mutate(newcol = (X2 - beta0here) ^ 2 ) %>%
  dplyr::summarize(rmse = sqrt(sum(newcol) / 1000))
beta1bias <- mean(result$V3)-beta1here
beta1RMSE <- result %>%
```

```
  dplyr::mutate(newcol1 = (V3 -beta1here) ^ 2 ) %>%
  dplyr::summarize(rmse1 = sqrt(sum(newcol1) / 1000))

summarize_result5 <- data.frame(matrix(nrow = 2, ncol = 2))
colnames(summarize_result5) <- c("beta0","beta1")
rownames(summarize_result5) <- c("Bias", "RMSE")
summarize_result5[1,1] <- beta0bias
summarize_result5[1,2] <- beta1bias
summarize_result5[2,1] <- beta0RMSE
summarize_result5[2,2] <- beta1RMSE

########################################
#Tolerance Limit simulation
## Part one
### This is an example of 95% tolerance limit with 95% CI with p=0.95
#### For 99% tolerance limit with 95% CI, p=0.99
mean1 <- 5.3322
mean2 <- 4.1934
mean3 <- 3.0546
mean4 <- 1.9158
points <- 12
LOQ <- 2
nrow <-48
each<-12
sd<-1
result1 <- data.frame(matrix(nrow = 10, ncol = 2))
colnames(result1) <- c("i", "up", "cov","ZPQ")
for (sim in 1:1000) {
  df <- matrix(, nrow = nrow, ncol = 3)
  df[,1] <- rep(1,nrow)
  df[,2] <- rep(c(6,12,18,24), each=each)
```

```
  set.seed(sim)
  data1 <-  rnorm(n=points,mean=mean1,sd=sd)
  set.seed(sim)
  data2 <-  rnorm(n=points,mean=mean2,sd=sd)
  set.seed(sim)
  data3 <-  rnorm(n=points,mean=mean3,sd=sd)
  set.seed(sim)
  data4 <-  rtruncnorm(n=points,mean=mean4,sd=sd)
  df[,3] <- c(data1,data2,data3,data4)
# Fit LCR
X <- df[,1:2] # X with 1 in 1st column
lm.1<-lm(df[,3]~ df[,2])
betas.1 <- as.matrix(coef(lm.1))
sigma.1 <- sigma(lm.1)
i<-1
while(T){
  df <-data.frame(df)
  df$mu.1 <- betas.1[1,1]+ df$X2 * betas.1[2,1]
  df$Y.1 <- ifelse (df$X3 > LOQ, df$X3,
                df$mu.1 - sigma.1 * (dnorm(((LOQ-df$mu.1)/sigma.1), 0,1))/(pnorm(((LOQ-
df$mu.1)/sigma.1), 0, 1)))
  lm.2<-lm(Y.1 ~ X2, data=df) # using detected y and estimated censored y
  betas.2 <- as.matrix(coef(lm.2))
  detect.df <- filter(df,X3 > LOQ)
  censored.df<-filter(df,X3 <= LOQ)
  sigma.square.2 <- 1/(nrow(df)) * (sum((detect.df$Y.1-detect.df$mu.1)^2)+ sigma.1^2 *
                                sum(1 - ((LOQ-censored.df$mu.1)/sigma.1)*
                                    (dnorm(((LOQ-censored.df$mu.1)/sigma.1), 0,1))/
                                    (pnorm(((LOQ-censored.df$mu.1)/sigma.1), 0, 1))))
  sigma.2 <- sqrt(sigma.square.2)
  if (sigma.2 - sigma.1 < 1e-6 & t(betas.2-betas.1)%*%(betas.2-betas.1) < 1e-6 | i>1000) break
```

```
  betas.1 <- betas.2
  sigma.1 <- sigma.2
  i<-i+1
  beta01<-betas.1[1,1]
  beta11<-betas.1[2,1]
  sigma11<-sigma.1
}
# Tolerence limit 95% 95%
# 1.Zp p quantile of N(0,1)
ZPresult <- data.frame(matrix(nrow = 1000, ncol = 2))
colnames(ZPresult) <- c("i", "Zp")
for (seed in 1:1000) {
  times = 1000
  p=0.95
  set.seed(seed)
  x <- rnorm(n=times,mean=0,sd=1)
  zp <- quantile(x, p)
  # 2. muhat*
  beta0here <- beta01
  beta1here <- beta11
  sigmahere <- sigma11
  Time <- 24
  muhat <- (beta0here + beta1here*Time - 3)/1
  sigmahat <- sigmahere/1
  # 3.Zp
  Zp <- (zp-muhat)/sigmahat
  ZPresult[seed, 1] <- seed
  ZPresult[seed, 2] <- Zp
}
# ZP (p,q), q=1-a
p=0.95
```

```
q=0.95
ZPQ <- quantile(ZPresult$Zp, q)
#coverage
set.seed(sim)
datahere <- rtruncnorm(n=points,mean=mean4,sd=sd)
numberup<-sum(datahere>ZPQ)
cov <- 1-(numberup/12)
result1[sim, 1] <- sim
result1[sim, 2] <- numberup
result1[sim, 3] <- cov
result1[sim, 4] <- ZPQ
}

## Part two
### This is an example of 95% tolerance limit with 95% CI with p=0.95
#### For 99% tolerance limit with 95% CI, p=0.99
mean1 <- 5.3322
mean2 <- 4.1934
mean3 <- 3.0546
mean4 <- 1.9158
points <- 12
LOQ <- 1
MRL<-2
nrow <-48
each<-12
sd<-1
resultwt95 <- data.frame(matrix(nrow = 10, ncol = 2))
colnames(resultwt95) <- c("i", "wt")
for (sim in 1:1000) {
  df <- matrix(, nrow = nrow, ncol = 3)
  df[,1] <- rep(1,nrow)
```

```
  df[,2] <- rep(c(6,12,18,24), each=each)
  set.seed(sim)
  data1 <-  rnorm(n=points,mean=mean1,sd=sd)
  set.seed(sim)
  data2 <-  rnorm(n=points,mean=mean2,sd=sd)
  set.seed(sim)
  data3 <-  rnorm(n=points,mean=mean3,sd=sd)
  set.seed(sim)
  data4 <-  rtruncnorm(n=points,mean=mean4,sd=sd)
  df[,3] <- c(data1,data2,data3,data4)
  # Fit LCR
X <- df[,1:2] # X with 1 in 1st column
lm.1<-lm(df[,3]~ df[,2])
betas.1 <- as.matrix(coef(lm.1))
sigma.1 <- sigma(lm.1)
i<-1
while(T){
  df <-data.frame(df)
  df$mu.1 <- betas.1[1,1]+ df$X2 * betas.1[2,1]
  df$Y.1 <- ifelse (df$X3 > LOQ, df$X3,
                df$mu.1 - sigma.1 * (dnorm(((LOQ-df$mu.1)/sigma.1), 0,1))/(pnorm(((LOQ-
df$mu.1)/sigma.1), 0, 1)))
  lm.2<-lm(Y.1 ~ X2, data=df) # using detected y and estimated censored y
  betas.2 <- as.matrix(coef(lm.2))
  detect.df <- filter(df,X3 > LOQ)
  censored.df<-filter(df,X3 <= LOQ)
  sigma.square.2 <- 1/(nrow(df)) * (sum((detect.df$Y.1-detect.df$mu.1)^2)+ sigma.1^2 *
                                sum(1 - ((LOQ-censored.df$mu.1)/sigma.1)*
                                    (dnorm(((LOQ-censored.df$mu.1)/sigma.1), 0,1))/
                                    (pnorm(((LOQ-censored.df$mu.1)/sigma.1), 0, 1))))
  sigma.2 <- sqrt(sigma.square.2)
```

```
if (sigma.2 - sigma.1 < 1e-6 & t(betas.2-betas.1)%*%(betas.2-betas.1) < 1e-6 | i>1000) break
betas.1 <- betas.2
sigma.1 <- sigma.2
i<-i+1
  beta01<-betas.1[1,1]
  beta11<-betas.1[2,1]
  sigma11<-sigma.1
}
# Tolerence limit 95% 95%
# 1.Zp p quantile of N(0,1)
ZPresult <- data.frame(matrix(nrow = 1000, ncol = 2))
colnames(ZPresult) <- c("i", "Zp")
for (seed in 1:1000) {
  times = 1000
  p=0.95
  set.seed(seed)
  x <- rnorm(n=times,mean=0,sd=1)
  zp <- quantile(x, p)
  # 2. muhat*
  beta0here <- beta01
  beta1here <- beta11
  sigmahere <- sigma11
  Time <- 24
  muhat <- (beta0here + beta1here*24 - 3)/1
  sigmahat <- sigmahere/1
  # 3.Zp
  Zp <- (zp-muhat)/sigmahat
  ZPresult[seed, 1] <- seed
  ZPresult[seed, 2] <- Zp
}
```

```
  # ZP (p,q), q=1-a
  p=0.95
  q=0.95
  ZPQ <- quantile(ZPresult$Zp, q)
  dfhere <- data.frame(matrix(nrow = 40, ncol = 2))
  dfhere$Time <- seq(1:40)
  dfhere$y <- beta0 + beta1* dfplot$Time
  dfhere$TL <- dfhere$y + ZPQ*0.95
  wt = (beta0-MRL)/-beta1here + ZPQ
  resultwt95[sim, 1] <- sim
  resultwt95[sim, 2] <- wt
}
mean95<- mean(resultwt95$wt)
sd95<- sd(resultwt95$wt)
resultwt95$truewt<-rep(25,1000)
resultwt95$diff<-resultwt95[,2]-resultwt95$truewt
over<-subset(resultwt95, resultwt95$diff>0)
same<-subset(resultwt95, resultwt95$diff==0)
under<-subset(resultwt95, resultwt95$diff<0)
overestimate<-sum(resultwt95$diff>=0)/1000
underestimate<-1-overestimate
oversd<-sd(over$diff)
undersd<-sd(under$diff)
undermean<-mean(under$diff)
overrmean<-mean(over$diff)
```