# Statistical Analysis of Stability of Repackaged Pharmaceutical Products

By

**Seungju Lee**

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Masters in Science

(Mathematics)

at the

**CALIFORNIA STATE UNIVERSITY - CHANNEL ISLANDS**

2018

© 2018

Seungju Lee

ALL RIGHTS RESERVED

APPROVED FOR THE MATHEMATICS PROGRAM

Dec 19/2018

Dr. Alona Kryshchenko, Thesis Advisor Date

Dec 19/2018

Dr. Jorge Garcia, Thesis Committee Date

19 Dec. 2018

Dr. Brian Sittinger, Thesis Committee Date

APPROVED FOR THE UNIVERSITY

1/4/19

Dr. Osman Özturgut Date
Dean, Extended University
AVP, International Programs

# Dedication

This thesis is dedicated to..

God, who gave me the strength and patience to finish this study,

my beloved boyfriend, Eric, who has always supported me and gives me so much love,

그리고 언제나 나의 가능성을 믿고 응원해 준 사랑하는 나의 엄마에게.

# Acknowledgements

I would like to thank some people who have inspired me, helped me, and challenged me during my master's degree studies at CSU Channel Islands. A few words of thanks can't express all my appreciation for them, but I would like to try my best. I am so thankful to have met:

* Dr. Alona Kryshchenko, my thesis advisor, whose passion and techniques in statistical analysis have inspired me so much. She has opened my eyes to the regression analysis world, introduced me to this project, and guided and advised me to make this study complete and well achieved.

* Dr. Jorge Garcia and Dr. Brian Sittinger, my thesis committee members, who are great professors and I couldn't have asked for more. Dr. Garcia has been always available when I needed help or advice. Dr. Sittinger's passion and strength in mathematics always inspire me and he is so approachable.

* Dr. Ivona Grzegorczyk, my department chair, who always provides the best support for the graduate students.

* Doug Blackburn, from GSMS, who has given me the opportunity to work on this cool project and valued my work.

* and my CI mathematics graduate program friends, who have worked with me in various ways and given the greatest attention to my studies and I.

Camarillo, California

December 21, 2018

# Abstract

Packaging/repackaging is essential to ensure pharmaceutical product's stability (i.e. efficacy and safety). Before processing to package/repackage products into container/closures, there must be evidence to verify its stability. The most common verification methods are real-time study and statistical analysis. Real-time study can be very accurate, but it is also time-consuming, which is inadequate for industry use. On the other hand, statistical analysis is appropriate for industry use since it can take short-term data and obtain the results quickly. In particular, we apply regression analysis on pharmaceutical product container/closure data along its decision-making framework to determine a container height 95% confidence interval and estimate the shelf life. Using a freeware program R, we are able to conduct the analysis more quickly and efficiently.

# Table of Contents


**1 Introduction** **4**

**2 Methods** **6**

2.1 The Proposed Framework . . . . . 6

2.2 Multi-variable Linear Regression Analysis . . . . . 10

2.3 Algebraic Approach to Solving Inequalities . . . . . 13

2.4 Container Height Estimation . . . . . 13

2.5 ANalysis of COVAriance (ANCOVA) . . . . . 15

**3 Data Analysis** **18**

3.1 C/C Configuration Analysis . . . . . 18

3.2 Assay & Dissolution Analysis . . . . . 26

**4 Example of Data Analysis** **29**

4.1 C/C Configuration Data Analysis . . . . . 29

4.2 Assay Data Analysis . . . . . 35

4.3 Dissolution Data Analysis . . . . . 39

5 Conclusion and Discussion 44

Bibliography 46

# Chapter 1

# Introduction

Packaging pharmaceutical products is an important part in pharmaceutical industry. Product stability, quality, potency and safety are affected by various characteristics, including packaging material and size. Mishra [9] claims that the role of packaging is not as simple as putting pills in a container for the product; in particular, it is indeed becoming more important both to the successful marketing of products and to the health and safety of patients as the industry evolves. Ensuring product stability before packaging pharmaceutical products concerns both manufacturers and repackagers. A repackager is a firm who sells pharmaceutical products that they bought from the manufacturer or distributor and repackaged under label different from that of the manufacturer [3]. In that case, the repackager must ensure the pharmaceutical product's quality and stability and comply to required qualification

for pharmaceuticals. According to Price [11], stability testing is conducted physically, chemically, or micro-biologically and measures how the quality of a drug product and its packaging vary with time under the influence of environmental factors. Price also states that the most proper approach of gathering data for stability verification is to conduct real-time stability studies. However, real-time studies are very much time-dependent; so it is suggested by the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) to conduct the analytic method to study stability of pharmaceutical product [6].

In Chapter 2, we first address the framework and statistical methods to verify product's stability support, and in Chapter 3, we analyze these methods theoretically. Lastly in Chapter 4, we practically analyze the results obtained after using these methods on pharmaceutical product container/closure stability data examples.

# Chapter 2

# Methods

## 2.1 The Proposed Framework

In this section, we propose a decision-making framework that is processed before repackaging pharmaceutical products to verify their stability supports. Following this framework, we evaluate stability of product with the impact of container/closure (C/C) type. Before Step 1, we firstly receive a manufacturer's stability reports for a product, which includes C/C configuration measurements and assay and dissolution records over time. Then, we define a C/C for the same product, and determine whether this product can be repackaged with using this type of C/C in Steps 1, 2 and 3 of the proposed framework depicted as in Figure 2.1.1.

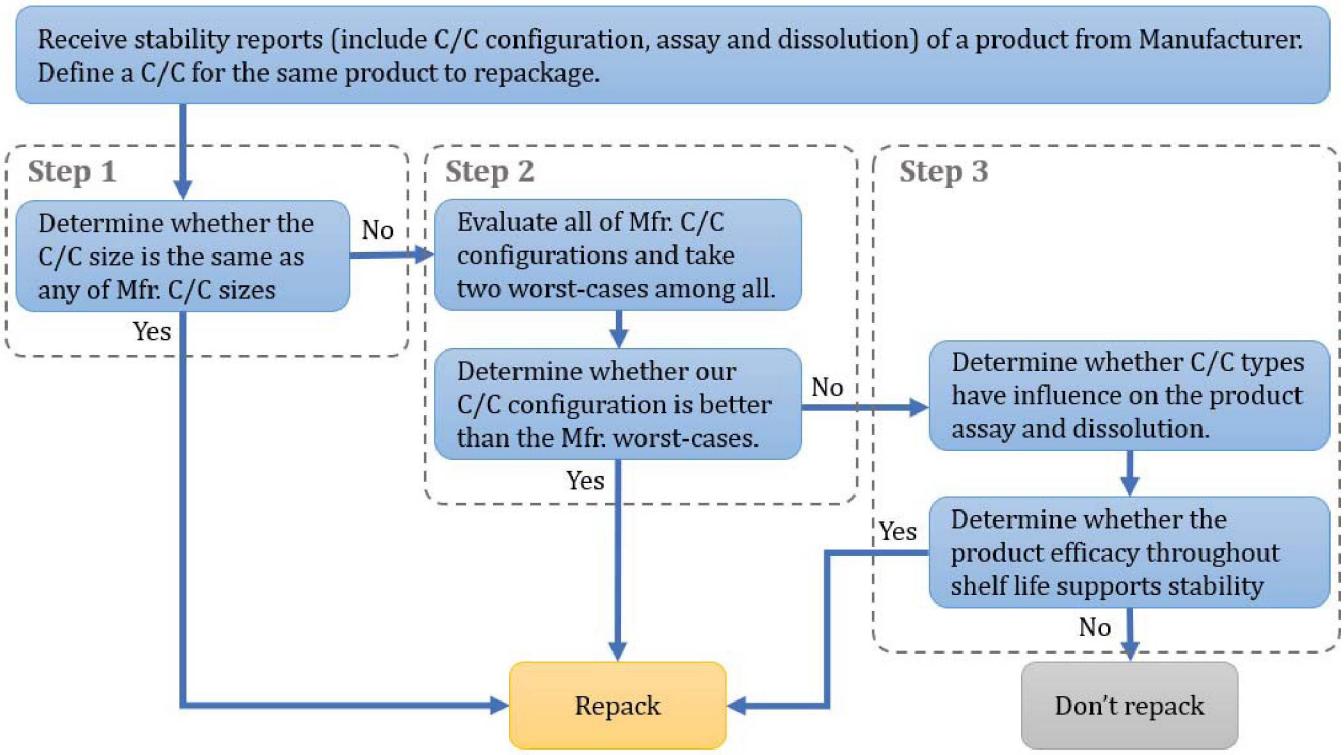


**Figure 2.1.1:** Decision-Making Frame Work

In Step 1, we check whether our C/C size is equivalent to any of manufacturer's C/C sizes, and if it is, then the product is repackaged with using that C/C. This is according to FDA guidance for industry which states that the packaging qualification information is not required if the repackager uses the same C/C approved in the original application or in the manufacturer's reports [3]. If a C/C has the equivalent size as what the manufacturer used for their packaging, then we can also use it for repackaging. However, if there is no equivalent C/C size from manufacturer's reports, then we proceed Step 2.

The "worst-cases" are two chosen C/C configurations from the manufacturer's reports,

which had the smallest container-surface-area to product-volume ratio or the largest headspace-volume to product-volume ratio among all; and each worst-case had one feature or the other. Denote these ratios of the worst-case features by $a$ and $b$, respectively; we detect $a$ and $b$ simply by computing each container-surface-area to product-volume ratio (CSA/PV) and headspace-volume to product-volume ratio (HV/PV) from the manufacturer's C/C configurations. The definitions of CSA/PV and HV/PV are described in detail in Section 3.1.3. In Figure 2.1.2, we use the given variables to classify the product space and headspace of a container. Then in Step 2, we verify whether our C/C has stable configuration for the product by comparing it with the worst-cases. The worst-cases are the bounds that our C/C must not exceed in order to guarantee its stability. We say that a C/C configuration is better than manufacturer's worst-cases whenever its CSA/PV is greater than $a$ and HV/PV is less than $b$. If it is better than the worst-cases, then the product is repackaged with using that C/C. However, if it did not lie within the worst-cases, then we proceed towards Step 3.

In Step 3, we firstly make sure that there are multiple batches that have been recorded for their assay values and dissolution rates. This is followed by using the Harmonised Tripartite Guideline for Evaluation for Stability Data [6], which state that the shelf life is estimated by statistical analysis of long-term data from at least three batches of a pharmaceutical product. Then we determine whether using different C/C types would affect on the product assay and dissolution. Based on the determination, the regression analysis to evaluate product efficacy

is performed differently. Note that product efficacy is indicated by the product's shelf life, which is the amount of time while a product will work both safely and effectively, and expressed in units of months [12]. In particular, a product's shelf life is the amount of time until its assay value does not drop below 90% [10] or it could be the amount of time until its dissolution rate does not drop below 80% (in 30 minutes) [5]. To estimate a product's shelf life, depending on the previous determination of impacts of C/C types, we will choose one from three possible regression models [8]. The product is repackaged after Step 3 if the shelf life is long enough to meet our customer's needs, and it is not repackaged if the shelf life is too short.

In general, most products are repacked after Step 2 before Step 3, and if some product with using a certain type of C/C is not qualified to be repacked after Step 3, then we consider over-label, pass it through or qualify a new C/C for the product.

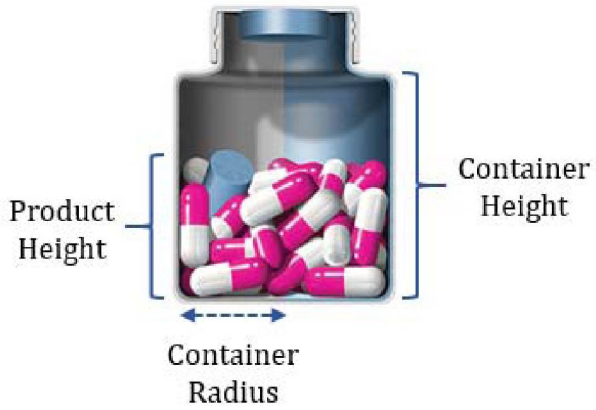


**Figure 2.1.2:** C/C Configuration Variables, the image adapted from [2]

## 2.2 Multi-variable Linear Regression Analysis

This approach follows from four *important questions* (1)-(4) that James et al. [7] addressed to proceed in order to confirm and improve a linear regression model:

(1) Is there any relationship between the response and the predictors of the model?

(2) Which of the predictors do certainly have impact on the response?

(3) How well does the model fit the data?

(4) Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Let $m$ denote the number of predictor variables. Then, for predictors $x_1, \cdots, x_m$ and corresponding coefficients $\beta_1, \cdots, \beta_m$, we set a regression model as

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m,$$

and address each of the four questions in turn.

(1) A relationship between product height and the predictors.

To determine whether there is a relationship between one variable ($Y$) and predictors, we use a hypothesis test. We test the null hypothesis,

$$H_0 \ : \ \beta_1 = \beta_2 = \cdots = \beta_m = 0, \tag{2.2.1}$$

versus the alternative hypothesis

$$H_a : \text{at least one } \beta_k \text{ is non-zero.} \tag{2.2.2}$$

Let $n$ denote the number of observed sample data points. For each $i = 1, \cdots, n$, let

$y_i$ denote the observation of $Y$ based on the $i^{th}$ values of each predictors,

$\hat{y}_i$ denote the prediction for $Y$ based on the $i^{th}$ values of each predictors, and

$\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

The residual sum of squares (RSS) and the total sum of squares (TSS) are defined as

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \quad \text{and} \quad \text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2.$$

The hypotheses (2.2.1) and (2.2.2) are tested by using the p-value approach and computing F-test statistic,

$$F = \frac{(\text{TSS} - \text{RSS})/m}{\text{RSS}/(n - m - 1)}. \tag{2.2.3}$$

If the p-value $< .05$, then we reject the null hypothesis, and we can conclude that there is a relationship between the variable ($Y$) and the predictors. If the p-value $\geq .05$, then we fail to reject the null hypothesis, and we conclude that the predictors have no relationship with the variable.

(2) Important predictor selection.

After (1), we must improve the regression model by keeping the predictors which

actually have impacts on the variable and eliminating the rest of the variables in the regression model. There are three methods from which we can choose to improve a regression model: forward, backward, and mixed selections. In this study, we use the backward selection. We start with all variables in the model as the predictors and compute each F-test statistics. If all variables have p-values $< .05$, then we keep them all in the model. If there is some variable with the p-value $> .05$, then we remove the variable with the largest p-value and conduct the regression analysis again for the regression model with one less predictor than before. We repeat removing variables until all remaining variables have p-values $< .05$.

(3) Model fit evaluation.

There are two common numerical measures of model fit: the relative standard error (RSE) and the square multiple correlation ($R^2$). Define $R^2$ by

$$R^2 = \frac{\sum_{i=1}^{n} |\hat{y}_i - \bar{y}|^2}{\text{TSS}}. \tag{2.2.4}$$

Observe that in fact, the numerator of $R^2$ is the regression sum of squares. Moreover, $R^2$ is always between 0 and 1, and the closer $R^2$ is to 1, the better the regression model fits the data.

(4) Predictions.

Since the prediction need not be exact, we use a 95% confidence interval of the model

to reflect the uncertainty.

## 2.3 Algebraic Approach to Solving Inequalities

This approach is to build one standard to compare with the worst-cases. In other words, instead of computing CSA/PV and HV/PV for every C/C system type and compare them with $a$ and $b$ (the worst-case ratios), we will compute the product height bounds that satisfy CSA/PV $> a$ and HV/PV $< b$. Note that both CSA/PV and HV/PV can be expressed by using $y, h$, and $r$. Since both ratios are in terms of $y$, the inequalities CSA/PV $> a$ and HV/PV $< b$ are solvable for $y$. Then by combining the two inequalities, we obtain the proper product height bounds. Hence, if the product height of a C/C configuration lies within the bounds, then this configuration is good to be used for repackaging. We use the product height bounds to combine with the product height regression model in the next section.

## 2.4 Container Height Estimation

We will obtain the 95% confidence interval of product height regression from Section 2.2 with $m$-many predictors as

$$y_L = L(\beta_{0,L}, \beta_{k,L}, x_k \ : \ k = 1, 2, \cdots, m) < y < U(\beta_{0,U}, \beta_{k,U}, x_k \ : \ k = 1, 2, \cdots, m) = y_U.$$

$y_L$ and $y_U$ denote the 95% lower and upper bounds of product height regression, respectively. Recall that we have the proper product height bounds from Section 2.3. Let $l(a, b, h, r)$ and $u(a, b, h, r)$ represent the lower and upper bounds of product height that are obtained by considering the worst-case boundaries. Then the proper product height bounds are expressed as

$$l(a, b, h, r) < y < u(a, b, h, r).$$

We combine these two bounds and obtain the following inequalities

$$l(a, b, h, r) < y_L \quad \text{and} \quad y_U < u(a, b, h, r). \tag{2.4.1}$$

In Chapter 3, we will see that the product height regression contains container volume variable as a predictor. Notice that the container volume is obtained by using container height $(h)$; so $y_L$ and $y_U$ contain $h$ in their expressions. This implies that if we solve the two inequalities $l(a, b, h, r) < y_L$ and $y_U < u(a, b, h, r)$ for $h$ and combine them, then we will obtain the 95% confidence interval for container height $(h)$.

In general, (2.4.1) depends on any of following variables: container height, container radius, and the regression predictors ($x_k$'s). In our study, we are particularly interested in finding the container height 95% confidence interval. For instance, if there are multiple containers with the same radius but different heights, then it would be convenient to compute the container height confidence interval for fixed values of container radius and other condi-

tions.

## 2.5 ANalysis of COVAriance (ANCOVA)

This approach facilitates estimating the shelf life by using statistical analysis that determines whether the variation is the same among different batches. Lee et al. suggest to use ANCOVA to add a covariance of time and batch to the regression [8]. Let

$A_i(t)$ denote the observed assay mean value of the $i^{th}$ batch at $t$ month,

$\alpha_0$ denote the common intercept of all batches,

$\alpha_i$ denote the deviation of the individual intercept of the $i$th batch,

$\beta_0$ denote the common slope of all batches,

$\beta_i$ denote the deviation of the individual slope of the $i^{th}$ batch, and

$\epsilon_{it}$ denote the model error term such that $\mathrm{E}(\epsilon_{it}) = 0$.

Then, we define the assay regression model as

$$A_i(t) = \alpha_0 + \alpha_i + (\beta_0 + \beta_i)t + \epsilon_{it}, \quad i = 1, \cdots, N, \tag{2.5.1}$$

where $N$ is the number of batches. There are two parts of hypothesis testing with using F-test statistic. One part is to test equality of intercepts (i.e. $H_{0,\alpha}$) and the other part is to test equality of slopes (i.e. $H_{0,\beta}$) of all batches which are expressed as

$$H_{0,\alpha} \ : \ \alpha_1 = \alpha_2 = \cdots = \alpha_N, \text{ and}$$

$$H_{0,\beta} \ : \ \beta_1 = \beta_2 = \cdots = \beta_N.$$

Then, the corresponding alternative hypotheses are

$$H_{a,\alpha} \ : \ \exists \ i, j \in \{1, \cdots, N\} \text{ such that } \alpha_i \neq \alpha_j, \text{ and}$$

$$H_{a,\beta} \ : \ \exists \ i, j \in \{1, \cdots, N\} \text{ such that } \beta_i \neq \beta_j.$$

Note that since U.S. Department of Health and Human Services Food and Drug Administration suggests to use a significance level of .25 [6]. Then, based on the p-values that are obtained from computing F-test statistics, there are three possible statistical results:

- $p_\alpha \geq 0.25$ and $p_\beta \geq 0.25$

  We fail to reject the null hypotheses at a significance level 0.25 for both testings. In this case, all batches from our data can be pooled, and the shelf life is estimated from the pooled data.

- $p_\alpha < 0.25$ and $p_\beta \geq 0.25$

  We reject the null hypothesis of equality of intercepts, and fail to reject the null hypothesis of equality of slopes. In this case, the data is pooled for all batches to estimate the common slope, but the intercepts may vary. Then, the shortest shelf life estimate among the batches is chosen to verify the stability support.

- $p_\alpha \geq 0.25$ and $p_\beta < 0.25$

  We reject the null hypothesis of equality of slopes, and fail to reject the null hypothesis of equality of intercepts. The shelf life for each batch is estimated individually without pooling, and the shortest estimate is chosen for stability support verification.

Similarly, the dissolution regression model is set as

$$D_i(t) = \alpha_0 + \alpha_i + (\beta_0 + \beta_i)t + \epsilon_{it}, \quad i = 1, \cdots, N, \tag{2.5.2}$$

where $D_i(t)$ denotes the observed dissolution mean value of the $i^{th}$ batch at $t$ month, and all other notations follow from the model (2.5.1). The testing hypotheses and test results also follow as the same as for assay regression model.

# Chapter 3

# Data Analysis

## 3.1 C/C Configuration Analysis

### 3.1.1 Data Description

A C/C configuration data set consists of 6 attributes as below:

- Strength (mg)

  The measured amount of the active ingredient present in each dosage [4].

- Pack Size (ct)

  The number of pills filled in the container.

- Container Type (Round/Square)

  "Round" for a round-type container, and "Square" for a square-type container.

- Container Radius/Side of Square (cm)

  The radius of the container bottom for a round-type container, and the length of one side of the container bottom for a square-type container.

- Container Height (cm)

  The height of the container.

- Product Height (cm)

  The shortest distance between the highest point of pills piled up inside the container and the bottom (see Figure 2.1.2).

Note that all variables above are numerical except for container type, which is binary. We also have manufacturer's C/C configuration data that contains the same attributes as the above C/C configuration data. Recall that the worst-cases are determined by obtaining the manufacturer's smallest CSA/PV and largest HV/PV (see Section 2.1). Notice that CSA/PV and HV/PV are computed by using the last three attributes above.

As stated above, there are round-type and square-type containers, but we will focus on analyzing round-type containers since analysis of square-type containers can be done similarly.

### 3.1.2 Product Height Regression Model

According to the attributes we have in this data, the predictor candidates are strength, pack size, container radius and container height. Considering that for containers with the same volume, their bottom radius and height could differ, we create a new variable, container volume. Let

$s_i$ denote the $i^{th}$ observed strength value

$p_i$ denote the $i^{th}$ observed pack size

$v_i$ denote the $i^{th}$ observed container volume

$\epsilon_i$ denote the model error term, which is normally distributed with the mean 0 and the standard deviation 1,

for each $i = 1, \cdots, n$. Let $\beta_s, \beta_p$, and $\beta_v$ be their corresponding coefficients and $\beta_0$ denote the regression intercept. Then the product height linear regression model is given as

$$y_i = \beta_0 + \beta_s s_i + \beta_p p_i + \beta_v v_i + \epsilon_i. \tag{3.1.1}$$

(1) Relationship between product height and the predictors - strength, pack size and container volume.

Using the parameters from the model (3.1.1), we test the null hypothesis

$$H_0 \ : \ \beta_s = \beta_p = \beta_v = 0$$

versus the alternative hypothesis

$$H_a : \text{at least one of } \beta_s, \beta_p \text{ and } \beta_v \text{ is not zero.}$$

And the hypotheses are tested by computing F-test statistic (2.2.3) with its p-value. If $p > .05$, then it implies that none of the three variables have relationship with product height, so we remove all of them from the regression model. If $p < .05$, then we keep all three variables $s, p$ and $v$.

(2) Important predictor selection.

All the variables that remain in the model after backward selection are the important predictors.

(3) Model fit evaluation.

Check $R^2$ for the regression model we obtain after (2). If the $R^2$ is close to 1, then this model is a good fit of the data, and if not, then it does not fit the data very well. There are two types of $R^2$: multiple $R^2$ and adjusted $R^2$. Multiple $R^2$ is $R^2$ for a multi-variable regression model, and adjusted $R^2$ is a modified version of multiple $R^2$ that has been adjusted for the number of predictors in the model [15]. We use the adjusted $R^2$ since multiple $R^2$ tends to increase whenever the number of variables increases in a model, and this means that multiple $R^2$ could be biased.

(4) Predictions.

For linear regression, the 95% confidence intervals for each regression parameters (regression intercept and coefficients) approximately take the form

$$\left[\hat{\beta}_k - 2 \cdot \mathrm{SE}(\hat{\beta}_k),\ \hat{\beta}_k + 2 \cdot \mathrm{SE}(\hat{\beta}_k)\right], \tag{3.1.2}$$

for each $k = 0, 1, \cdots, m$. Each upper and lower bounds of regression parameters will be applied in Section 3.1.4.

### 3.1.3 Algebraic Analysis

Denote C/C configurations, container radius, container height and product height as $r, h$ and $y$, respectively (see Figure 3.1.1). Then it follows that headspace height is $h - y$. Then CSA/PV and HV/PV are expressed as

$$\frac{\mathrm{CSA}}{\mathrm{PV}} = \frac{2\pi r^2 + 2\pi r h}{\pi r^2 y}, \quad \text{and} \quad \frac{\mathrm{HV}}{\mathrm{PV}} = \frac{\pi r^2 (h - y)}{\pi r^2 y}. \tag{3.1.3}$$

Note that $r$ and $h$ are fixed for a C/C type, but $y$ may vary within the same C/C type for each configuration. Recall that in order to use a C/C for repacking, it must have a better configuration than the manufacturer's two worst-cases which have $a$, the smallest CSA/PV, and $b$, the largest HV/PV. In other words, it must satisfy $\frac{\mathrm{CSA}}{\mathrm{PV}} > a$ and $\frac{\mathrm{HV}}{\mathrm{PV}} < b$. Suppose computing CSA/PV and HV/PV for each C/C configuration to compare with the manufacturer's worst-cases. For a large size of data, we realize that it is not efficient to do so. Instead, we analyze the worst-case ratios, and the worst-case criteria (3.1.3) can be set

in terms of product height. Consider

$$\frac{\text{CSA}}{\text{PV}} = \frac{2\pi r^2 + 2\pi rh}{\pi r^2 y} > a \tag{3.1.4}$$

and

$$\frac{\text{HV}}{\text{PV}} = \frac{\pi r^2 (h - y)}{\pi r^2 y} < b. \tag{3.1.5}$$

Now, modify these two inequalities (3.1.4) and (3.1.5) to express $y$. Then we obtain

$$\frac{h}{b+1} < y < \frac{2r + 2h}{ar}. \tag{3.1.6}$$

We will call (3.1.6) the proper product height bounds, and it indicates what product height values are qualified to be repacked.

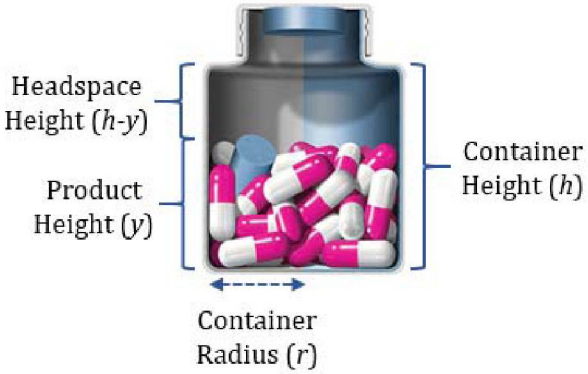


**Figure 3.1.1:** Definition of Configuration Attributes, the image adapted from [2]

### 3.1.4 Container Height Analysis

By combining the proper product height bounds (3.1.6) and the product height regression model (3.1.1), we obtain

$$\frac{h}{b+1} < \beta_0 + \beta_s s + \beta_p p + \beta_v v < \frac{2r+2h}{ar}. \tag{3.1.7}$$

Note that if any predictor from the regression model (3.1.1) had been removed, then its parameter will perform as zero in the bounds (3.1.7).

We will use the upper and lower bound expressions of the 95% confidence intervals of the regression parameters as obtained in (3.1.2). Denote the each expressions as $\beta_{k,a}$ and $\beta_{k,b}$, respectively. Then, we apply $\beta_{k,a}$ and $\beta_{k,b}$ to the inequality (3.1.7), then we obtain

$$\beta_{0,a} + \beta_{s,a} s + \beta_{p,a} p + \beta_{v,a} \pi r^2 h < \frac{2r+2h}{ar}, \tag{3.1.8}$$

and

$$\frac{h}{b+1} < \beta_{0,b} + \beta_{s,b} s + \beta_{p,b} p + \beta_{v,b} \pi r^2 h. \tag{3.1.9}$$

Then, after solving the inequality (3.1.8) for $h$, we obtain

$$h > B_1, \quad \text{if} \quad \beta_{v,a} < \frac{2}{ar^3}, \tag{3.1.10}$$

and

$$h < B_1, \quad \text{if} \quad \beta_{v,a} > \frac{2}{ar^3}, \tag{3.1.11}$$

where

$$B_1 = \frac{ar(\beta_{0,a} + \beta_{s,a}s + \beta_{p,a}p) - 2r}{2 - \pi\beta_{v,a}ar^3}. \tag{3.1.12}$$

Similarly, after solving the inequality (3.1.9) for $h$, we get

$$h < B_2, \quad \text{if} \quad \beta_{v,b} < \frac{1}{(b+1)\pi r^2}, \tag{3.1.13}$$

and

$$h > B_2, \quad \text{if} \quad \beta_{v,b} > \frac{1}{(b+1)\pi r^2}, \tag{3.1.14}$$

where

$$B_2 = \frac{(b+1)(\beta_{0,b} + \beta_{s,b}s + \beta_{p,b}p)}{1 - (b+1)\beta_{v,b}\pi r^2}. \tag{3.1.15}$$

There are four scenarios of container height 95% CI depending on the values obtained for $\beta_{v,a}$ and $\beta_{v,b}$:

(i) $(B_1, B_2)$, if $\beta_{v,a} < \frac{2}{ar^3}$ and $\beta_{v,b} < \frac{1}{(b+1)\pi r^2}$ : combined (3.1.10) and (3.1.13)

(ii) $(B_2, B_1)$, if $\beta_{v,a} > \frac{2}{ar^3}$ and $\beta_{v,b} > \frac{1}{(b+1)\pi r^2}$ : combined (3.1.11) and (3.1.14)

(iii) $(0, \min(B_1, B_2))$, if $\beta_{v,a} > \frac{2}{ar^3}$ and $\beta_{v,b} < \frac{1}{(b+1)\pi r^2}$ : combined (3.1.11) and (3.1.13)

(iv) $(\max(B_1, B_2), +\infty)$, if $\beta_{v,a} < \frac{2}{ar^3}$ and $\beta_{v,b} > \frac{1}{(b+1)\pi r^2}$ : combined (3.1.10) and (3.1.14)

In fact, the scenario (i) happens in most cases of C/C configurations.

## 3.2 Assay & Dissolution Analysis

### 3.2.1 Data Description

An assay & dissolution data set consists of 4 attributes as below:

- Batch
- Time (months)
- Assay (%)
- Dissolution Rate (%)

The Batch is a characteristic variable, which indicates labeled names of each batch. The Time variable indicates each time point when each assay and dissolution were measured. The Assay represents the purity value of a drug substance tested by using one of physical, colormetric and analytic methods [14]. The Dissolution represents extent of drug absorption in 30 minutes [13].

### 3.2.2 Regression Analysis with ANCOVA

Recall that analysis on assay & dissolution data is required whenever a C/C sample did not qualify to be repacked after Step 2. In Step 3, the shelf life of a product is estimated using ANCOVA and is used to ensure that the shelf life meets customers' needs. The shelf life from

assay testing is the first time when its assay drops below 90%, and shelf life from dissolution testing occurs when its dissolution rate drops below 80%. Based on the ANCOVA results, we have three scenarios for assay estimation; and dissolution estimation is performed in the same way.

- $p_\alpha \geq 0.25$ and $p_\beta \geq 0.25$

  Since all batches are pooled, we perform the model

  $$A(t) = \alpha + \beta t. \tag{3.2.1}$$

  The shelf life is estimated by using the 95% confidence lower bounds of the regression intercept and coefficient.

- $p_\alpha < 0.25$ and $p_\beta \geq 0.25$

  Since batches have the same slope but different the intercepts, we perform regression models for all $N$-many batches, which have the form

  $$A_i(t) = \alpha_0 + \alpha_i + \beta t, \;\; i = 1, \cdots, N \tag{3.2.2}$$

  where $\alpha_0 + \alpha_i$ denotes the intercept for the $i^{th}$ batch. Then, among all batches, there are at least two batches with distinct regression intercepts.

  Then we obtain up to $N$-many parallel regression models, which means that the model with the smallest intercept will result in the shortest shelf life. Thus, the shelf life

estimation is required to use the smallest value among the 95% confidence lower bounds of all distinct model intercepts and the 95% confidence lower bound of the common model slope.

- $p_\alpha < 0.25$ and $p_\beta < 0.25$

  In this scenario, we obtain $N$-many regression models,

$$A_i(t) = \alpha_0 + \alpha_i + (\beta_0 + \beta_i)t, \;\; i = 1, \cdots, N \tag{3.2.3}$$

  where $\alpha_0 + \alpha_i$ and $\beta_0 + \beta_i$ denote the intercept and the slope for the $i^{th}$ batch, respectively. In addition to the condition in the previous scenario, we also have that among all batches, there are at least two batches with distinct regression slopes.

  Then we take the 95% confidence lower bounds of each regression slopes and intercepts and estimate a shelf life individually. The shortest estimate among all will be chosen to be the shelf life.

For each aforementioned scenario, let $\alpha^*$ and $\beta^*$ denote the regression intercept and slope such that for each $i = 1, \cdots, N$,

$$\alpha^* = \alpha_0 + \alpha_i \quad \text{and} \quad \beta^* = \beta_0 + \beta_i.$$

Then, the 95% confidence lower bounds for $\alpha^*$ and $\beta^*$ are obtained as

$$\left[\hat{\alpha^*} - 2 \cdot \text{SE}(\hat{\alpha^*}),\ \hat{\alpha^*} + 2 \cdot \text{SE}(\hat{\alpha^*})\right] \quad \text{and} \quad \left[\hat{\beta^*} - 2 \cdot \text{SE}(\hat{\beta^*}),\ \hat{\beta^*} + 2 \cdot \text{SE}(\hat{\beta^*})\right].$$

# Chapter 4

# Example of Data Analysis

Since the aim of this study is to conduct statistical analysis for industry use, we analyze the our sample data using the freeware program R. This program is often used in statistics, and can be employed in a user-friendly manner if it is converted properly to R-Markdown.

## 4.1 C/C Configuration Data Analysis

(1) Relationship between product height, and the predictors - strength, pack size and container volume.

We perform regression analysis on the regression model

$$y = \beta_0 + \beta_s s + \beta_p p + \beta_v v. \tag{4.1.1}$$

Then the analysis summary is obtained as in R output 4.1. Since we obtained the p-value for F-statistic (on the bottom of R output 4.1) is $5.968e-13 < .05$, we conclude that a relationship between product height and the three predictor variables has been detected.

**R output 4.1:** Product Height Regression Model With All Predictor Candidates

```
Call:
lm(formula = Product.Height ~ Strength + Pack.Size +
   Container.Volume, data = stabilityData)


Residuals:
     Min       1Q   Median       3Q      Max
-0.37313 -0.09915  0.00184  0.08641  0.33990


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.551982   0.171737   9.037 7.37e-09
Strength         -0.015173   0.061194  -0.248    0.806
Pack.Size         0.015328   0.001010  15.171 3.91e-13
Container.Volume -0.015858   0.001658  -9.567 2.68e-09
---
```

```
Residual standard error: 0.1753 on 22 degrees of freedom
Multiple R-squared:  0.9314,    Adjusted R-squared:  0.922
F-statistic:  99.5 on 3 and 22 DF,  p-value: 5.968e-13
```

(2) Important predictor selection.

Notice that in R output 4.1, the p-values we obtain for each variable are obtained from t-test statistic. Since $t^2$ and F-statistics are equivalent [1], they result in the same p-values. Hence, the p-values are valid to use for backward selection. Note that p-value for strength variable is .806 > .05, so eliminate strength variable from the regression and perform the analysis again (see R output 4.2). The model after one elimination has all its variables with p-values < .05, which indicates that the variables, pack size, and container volume are important predictors of the model.

**R output 4.2:** Product Height Regression Model After One Elimination

```
Call:
lm(formula = Product.Height ~ Pack.Size + Container.Volume,
   data = stabilityData)


Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.38867 -0.10659  0.00088  0.09446  0.32698


Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.5300180  0.1440958  10.618 2.44e-10
Pack.Size         0.0153484  0.0009864  15.560 1.06e-13
Container.Volume -0.0157977  0.0016059  -9.838 1.04e-09
---
Residual standard error: 0.1717 on 23 degrees of freedom
Multiple R-squared:  0.9312,    Adjusted R-squared:  0.9252
F-statistic: 155.6 on 2 and 23 DF,  p-value: 4.312e-14
```

(3) Model fit evaluation.

In R output 4.2, notice that the adjusted $R^2$ is obtained as $R^2 = .9252$, which is relatively close to 1. This implies that the predicted product height values by using this regression model fits well with the observed product height values. The model's fitness with the actual data is also depicted in Figure 4.1.1. The horizontal axis represents observed product height values, and the vertical axis represents predicted product height values based on the regression model. The closer the points are to the blue line ($y = x$), the more predicted values match the observed values.

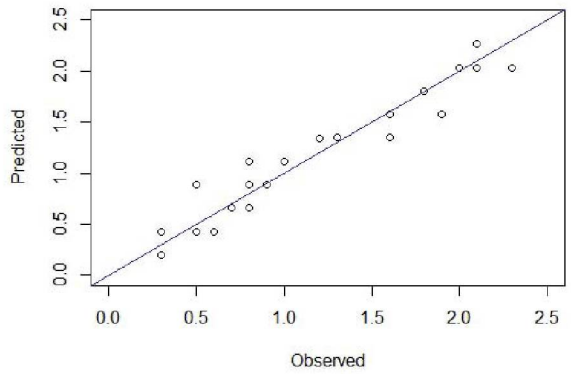


**Figure 4.1.1:** Product Height (cm) : Observed vs. Predicted ($R^2 = .9252$)

(4) Predictions.

The upper and lower bounds of each parameters are used to compute container height 95% confidence interval. We obtain

- $\beta_{0,a} \approx 1.2319$
- $\beta_{0,b} \approx 1.8281$
- $\beta_{p,a} \approx .0133$
- $\beta_{p,b} \approx .0174$
- $\beta_{v,a} \approx -.0191$
- $\beta_{v,b} \approx -.0125$

as in R output 4.3.

**R output 4.3:** The 95% confidence upper and lower bounds of each parameters

```
                    2.5 %       97.5 %
(Intercept)      1.23193319   1.82810275
```

```
Pack.Size          0.01330793   0.01738894

Container.Volume  -0.01911963  -0.01247568
```

We will use the above parameters to estimate container height 95% confidence interval that is qualified to be repacked.

Suppose that manufacturer's worst-case ratios for some product are $a = 1.16$ and $b = 2.31$ and we want to pack 60 pills in a container of the bottom radius 2cm. Then, first we check the conditions for $\beta_{v,a} = -0.01247568$ and $\beta_{v,b} = -0.01911963$.

$$\frac{2}{ar^3} = \frac{2}{1.16 \cdot 2^3} \approx 0.21 > -0.01247568 = \beta_{v,a}$$
$$\frac{1}{(b+1)\pi r^2} = \frac{1}{3.31 \cdot \pi \cdot 2^2} \approx 0.02 > -0.01911963 = \beta_{v,b}$$

Notice that this is the scenario (i) from 3.1.4. Then the 95% confidence interval for container height (cm) is obtained as

$$\left( \frac{ar(\beta_{0,a} + \beta_{p,a}p) - 2r}{2 - \pi\beta_{v,a}ar^3}, \frac{(b+1)(\beta_{0,b} + \beta_{p,b}p)}{1 - (b+1)\beta_{v,b}\pi r^2} \right) \approx (0.2778387, 6.257375).$$

Hence, to repack 60 pills in a C/C with its bottom radius 2 cm with stability supports, we must make sure that the container height is in between .2778 cm and 6.2574 cm with 95% confidence.

## 4.2 Assay Data Analysis

To set up a regression model to predict assay, we firstly determine whether different batches are to be considered separate or pooled. This can be done by using ANCOVA on assay values over covariance of time and batch. Consider R output 4.4. F-test statistic p-value for Batch (i.e. $p_\alpha = .02067$) and for Batch:Time (i.e. $p_\beta = .04893$) are both smaller than .25. This implies that the intercept and the slope are both pooled, so there will be one regression model that covers all batches.

**R output 4.4:** ANCOVA: Batch*Time vs. Assay (%)

```
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
Batch        4   22.30    5.576    3.581  0.02067
Time         1   17.89   17.893   11.493  0.00252
Batch:Time   4   17.53    4.381    2.814  0.04893
Residuals   23   35.81    1.557
```

**R output 4.5:** Linear Regression Model: Time vs. Assay (%)

```
Call:
lm(formula = Assay ~ Batch * Time, data = stabilityData)


Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2.37635 -0.74192 -0.02006  0.92036  1.76737


Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      98.560479   0.778430 126.615  < 2e-16
BatchBatch2       1.623952   1.100866   1.475 0.153729
BatchBatch3       2.129940   1.100866   1.935 0.065409
BatchBatch4       4.350299   1.100866   3.952 0.000634
BatchBatch5      -0.540479   1.240988  -0.436 0.667243
Time             -0.004491   0.060211  -0.075 0.941187
BatchBatch2:Time -0.074551   0.085151  -0.876 0.390347
BatchBatch3:Time -0.089022   0.085151  -1.045 0.306675
BatchBatch4:Time -0.254890   0.085151  -2.993 0.006490
BatchBatch5:Time -0.010509   0.089162  -0.118 0.907199
---


Residual standard error: 1.248 on 23 degrees of freedom
Multiple R-squared:  0.6172,    Adjusted R-squared:  0.4673
F-statistic:  4.12 on 9 and 23 DF,  p-value: 0.002905
```

Then, according to the regression model form (3.2.3) from Section 3.2.2, by using the coefficients from R output 4.5, we set the regression models as following:

$$(\text{Assay}_1) = 98.560479 - .004491 \times (\text{Time}),$$

$$(\text{Assay}_2) = (98.560479 + 1.623952) - (.004491 + .074551) \times (\text{Time}),$$

$$(\text{Assay}_3) = (98.560479 + 2.129940) - (.004491 + .089022) \times (\text{Time}),$$

$$(\text{Assay}_4) = (98.560479 + 4.350299) - (.004491 + .254890) \times (\text{Time}),$$

$$(\text{Assay}_5) = (98.560479 - .540479) - (.004491 + .010509) \times (\text{Time}).$$

The Figure 4.2.1 describes the observed assay values as black dots, the 5 regression fits as solid lines, and the 95% lower bounds of the regressions as dashed lines. The colors of solid and dashed lines match for the same batch. The shelf life of the product is chosen as the earliest time when the regression 95% lower bounds hit the assay low specification limit (= 90%). This leads us to

$$90 = \alpha^* + \beta^* \times (\text{Time}),$$

and solving for (Time) yields

$$(\text{Time}) = \frac{90 - \alpha^*}{\beta^*}. \tag{4.2.1}$$

By using the formula 4.2.1, we have 5 shelf life estimates, and the smallest value among the 5 estimates is chosen to be the shelf life of this product. In fact, the estimated shelf life is

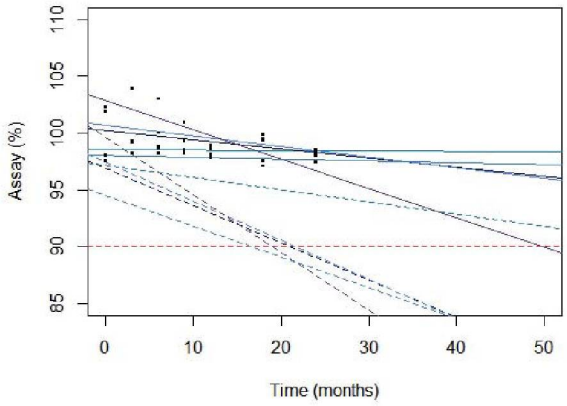


**Figure 4.2.1:** Shelf Life Estimation = 16.8 months (Assay over 100% occurs due to measurement uncertainty in laboratory reports.)

16.8 months with 95% confidence. From R output 4.6, we obtain the 95% lower bounds of each regression intercepts and slopes.

**R output 4.6:** 95% Lower and Upper Bounds of Assay Regression Coefficients

```
                        5 %         95 %
(Intercept)       97.2263508 99.89460733
BatchBatch2       -0.2627902  3.51069441
BatchBatch3        0.2431978  4.01668244
BatchBatch4        2.4635571  6.23704172
BatchBatch5       -2.6673735  1.58641539
Time              -0.1076848  0.09870280
BatchBatch2:Time  -0.2204890  0.07138720
```

```
BatchBatch3:Time  -0.2349601   0.05691614

BatchBatch4:Time  -0.4008283  -0.10895212

BatchBatch5:Time  -0.1633215   0.14230351
```

If a customer needs the product for at least 30 months of use, then we conclude that this product cannot be repackaged since the estimated shelf life based on assay analysis is not enough to fulfill the customer's need.

## 4.3 Dissolution Data Analysis

Similarly as in Section 4.2, we must determine whether different batches are to be considered separate or pooled in order to set a regression model to predict dissolution rates. By using ANCOVA on dissolution rates over covariance of time and batch, we obtain as in R output 4.7. The F-test statistic p-value for Batch (i.e. $p_\alpha = .552$) and for Batch:Time (i.e. $p_\beta = .468$) are both greater than .25. This implies that the intercept and the slope are both not to be pooled, so we need to perform regression models individually for each batches (i.e. the first scenario (3.2.1) from Section 3.2.2).

**R output 4.7:** ANCOVA: Batch*Time vs. Dissolution (% in 30 min)

```
             Df  Sum Sq  Mean Sq  F value  Pr(>F)
Batch         4   12.96    3.241    0.775   0.552
Time          1    1.38    1.381    0.330   0.571
Batch:Time    4   15.41    3.852    0.922   0.468
Residuals    23   96.13    4.179
```

Then, as we have analyzed in Section 3.2.2, we set the regression model as

$$(\text{Dissolution}) = 00.70720 - .02971 \times (\text{Time}),$$

by using the coefficients obtained as in R output 4.8. Figure 4.3.1 describes the observed dissolution rates as black dots, the regression fit as a green solid line, and the 95% lower bound of the regression as a green dashed line. The shelf life of the product is defined to be the time when the regression 95% lower bound hits the dissolution low specification limit ($= 80\%$).

**R output 4.8:** Linear Regression Model: Time vs. Assay (%)

```
Call:
lm(formula = Dissolution ~ Time, data = stabilityData)


Residuals:
```

```
    Min       1Q   Median       3Q      Max

-3.4398  -1.3507  -0.1725   1.2928   3.4710


Coefficients:

             Estimate  Std. Error  t value  Pr(>|t|)

(Intercept) 100.70720     0.57803  174.224    <2e-16

Time          -0.02971     0.04375   -0.679     0.502

---


Residual standard error: 2 on 31 degrees of freedom

Multiple R-squared:  0.01465,    Adjusted R-squared:  -0.01713

F-statistic: 0.461 on 1 and 31 DF,  p-value: 0.5022
```

The estimated shelf life is obtained as 189.9 months with 95% confidence algebraically, by using the lower bounds of the regression coefficients (see R output 4.9):

$$80 = 99.7271374 - .1038883 \times (\text{Time}),$$

$$(\text{Time}) = \frac{99.7271374 - 80}{.1038883} \approx 189.9.$$

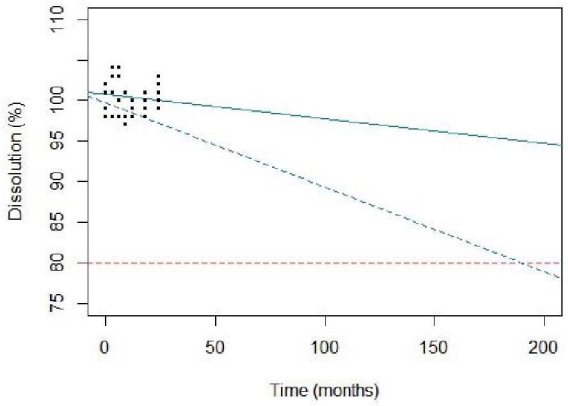


**Figure 4.3.1:** Shelf Life Estimation = 189.9 months (Dissolution over 100% occurs due to measurement uncertainty in laboratory reports.)

**R output 4.9:** 95% Lower and Upper Bounds of Dissolution Regression Coefficients

```
                  5 %          95 %
(Intercept) 99.7271374 101.68726761
Time         -0.1038883   0.04447638
```

In fact, the sample data we have used for dissolution analysis in this section is the same as the sample data we used for assay analysis in Section 4.2. Recall that in Section 4.2, the shelf life estimate was 16.8 months. Note that shelf life must be the earliest estimate of the product dropping below the low specification. Hence, in this case, the shelf life of the sample product is determined to be 16.8 months rather than 189.9 months. In other words, if we obtain the product's shelf life estimate as two different amounts of time between assay

and dissolution analysis, then the shorter estimate is chosen to be the practical shelf life of the product. Thus, if a customer needs the product for at least 30 months of use, then we conclude with 95% confidence that this product should not be repacked since the shelf life is estimated to be 16.8 months, which is not enough to fulfill the customer's need. On the other hand, if a customer needs the product for up to 1 year, then this product can be repacked with 95% confidence in its stability support, since the shelf life estimate is longer than the customer's need.

# Chapter 5

# Conclusion and Discussion

Pharmaceutical repackaging include additional steps before processing to verify its stability supports from regular packaging. These additional steps are indicated as Step **1** and **2** in Figure 2.1.1, and these steps do not cause more work to do. They actually reduce the work during the process. If products are repacked after Step **1** or **2**, then Step **3** is skipped which is the most time-dependant. In fact, Step **3** does not take much place in decision-making process before repackaging. Most products are repackaged after Step **2**. Step **2** is more flexible since if a container/closure configuration did not qualify to be repacked, then it can be just reconfigured, such as by changing the pack size or the container.

There are some limitations of the methods used for Step **2** and Step **3** in this study. Lee et al. [8] stated that Step **3** method assumes that assay has a linear regression, even though

it is not known whether it would be linear. Step **2** method also uses linear regression analysis for product height, and it is not known whether it would be linear. Hence, the results could be biased if they were actually not linear.

This study saves time on performing Step **2** and **3**, especially with using R, and it absolutely results in labor-saving. In fact, Step **2** became simplified from computing individual configurations CSA/PV and HV/PV ratios and comparing individually to computing container height 95% confidence interval and configuring the appropriate container/closures. Moreover, Step **3** became capable to process without measuring assay and dissolution for long amount of time, such as 50 months. Instead, by using about 24-month of data, the shelf life can be estimated statistically.

# Bibliography


[1] D. D. Boos and L. A. Stefanski, *Essential Statistical Inference: Theory and Methods*, Springer, 2013.

[2] Clariant, Healthcare Packaging, *Enhancing and Protecting Healthcare Products Through Polymer and Desiccant Packaging Solutions*, 2016 (accessed December 19, 2018). https://www.clariant.com/en/Business-Units/Masterbatches/Healthcare-Medical-and-Pharmaceutical/Healthcare-Packaging.

[3] FDA, *Guidance for industry container closure systems for packaging human drugs and biologic*, (1999).

[4] FDA, *Glossary of Terms*, November 2017 (accessed October 23, 2018). https://www.fda.gov/drugs/informationondrugs/ucm079436.htm.

[5] FDA, *Dissolution Testing and Acceptance Criteria for Immediate-Release Solid Oral Dosage Form Drug Products Containing High Solubility Drug Substances*,

March 2018 (accessed November 12, 2018). http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm.

[6] ICH, *Q1e harmonised tripartite guideline for evaluation for stability data*, (2004).

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer, 2013.

[8] H. Lee, P. Wu, and Y. Lee, *stab: An r package for drug stability data analysis*, Computer Methods and Programs in Biomedicine, 100 (2010), pp. 140–148.

[9] S. Mishra, *Rethinking The Role Of Packaging Design In Drug Development*, June 2018 (accessed November 12, 2018). https://www.pharmaceuticalonline.com/doc/rethinking-the-role-of-packaging-design-in-drug-development-0001.

[10] K. Naveršnik and R. Jurečič, *Humidity-corrected arrhenius equation: The reference condition approach*, International Journal of Pharmaceutics, 500 (2016), pp. 360–365.

[11] E. Price, *How to Know when to Toss your Prescription Drug or Refrigerate it*, June 2017 (accessed October 17, 2018). https://www.pcisynthesis.com/how-to-know-when-to-toss-your-prescription-drug-or-refrigerate-it/.

[12] M. PUGLIELLI, *What Is Shelf Life Testing Of A Pharmaceutical Product?*, November 2014 (accessed November 7, 2018). https://www.wellspringcmo.com/blog/what-is-shelf-life-testing-of-pharmaceutical-product.

[13] A. SIEW, *Dissolution testing*, 40 (2016), pp. 56,64.

[14] J. SUICO, *How Can We Check the Purity of a Substance?*, May 2018 (accessed November 12, 2018). https://sciencing.com/can-check-purity-substance-5950683.html.

[15] R. WOODHOUSE, *Statistical Regression Line-Fitting in the Oil and Gas Industry*, PennWell Corporation, 2003.