BSTT562 Project I

Ruizhe Chen, Hesen Li

December 14th, 2018

1 Abstract

There has been a long existing problem in analytical chemistry: handling uncertainty in chemical measurements. Major challenges in this type of problem are heteroscedasticity, which results from increasing variances with increasing concentrations and inter-laboratories variations. Acknowledging the fact that uncertainty in measurements should be incorporated in computing bounds on the underlying true concentration, Bhaumik and Gibbons [2005] proposed a random-effect calibration curve model. They used data separation technique in computing the point estimates and confidence intervals for true unknown concentrations X that varies from low to high levels from multiple laboratories. In this final project, we want to reproduce their analysis procedure for a real data set that is presented in (Bhaumik and Gibbons [2005]) using their proposed methods.

2 Data

To illustrate the properties of the point estimate and confidence region for the true concentration X, Bhaumik and Gibbons [2005] analyzed experimental data for cadmium from an interlaboratory study conducted by the Ford Motor Company (J. Phillips, personal communication). These data were generated as part of a blind interlaboratory study of laboratories that hold Michigan State Drinking Water Certifications for the parameters tested. The data has the characteristic of heteroscedasticity such that measurements at higher concentration levels have greater variability than those at lower levels.

The samples were prepared by an independent source, randomized, and submitted on a weekly basis over a 5-week period. Cadmium was analyzed by inductively coupled plasma atomic emissions spectroscopy using EPA method 200.7. The data set used for this example comprised five replicates at each of three concentrations (0, 20 and 100 $\mu g/L$) in each of the $q = 5$ laboratories. Using the first replicate from the first three laboratories as the new measurement (i.e., $q' = 3$), we would like to reproduce the results of point estimates, variances, confidence intervals and simulated confidence levels for the true concentrations in each of the three cases. The data is displayed as follows:

Table 1: Interlaboratory Data for Cadmium (ug / L)

Lab	Replication	0	20	100
1	1	-3.000	10.00	92.00

Lab	Replication	0	20	100
1	2	4.000	20.00	100.00
1	3	-4.000	17.20	97.80
1	4	3.000	24.00	100.00
1	5	3.100	19.10	109.00
2	1	-0.060	17.82	90.45
2	2	0.010	17.30	87.61
2	3	0.115	16.57	85.55
2	4	-0.055	17.36	89.92
2	5	0.340	18.12	90.07
3	1	-7.400	27.10	107.40
3	2	-2.100	19.40	108.10
3	3	-11.400	9.00	83.80
3	4	-11.100	10.50	81.90
3	5	-1.400	19.30	94.20
4	1	1.000	21.00	96.00
4	2	-2.126	16.05	90.65
4	3	0.523	16.08	89.39
4	4	-2.000	17.00	91.00
4	5	-0.551	15.49	85.87
5	1	0.000	18.00	91.00
5	2	0.000	19.00	101.00
5	3	0.000	19.00	102.00
5	4	-1.000	18.70	92.70
5	5	0.038	19.79	99.88

3 Model

To measure the true concentration of an analyte, x, a simple and straightforward method is to propose a linear calibration model as:

$$y = \alpha + \beta x + e \tag{1}$$

with normal assumption on errors. However because our data set has heteroscedastic variances, this simple linear models fails to explain the increasing measurement variation with increasing analyte concentration, which is also commonly observed in analytic data. On the other hand, a log-linear model, for example, $y = xe^{\eta}$, where η is a normal random variable with mean 0 and standard deviation σ_η also fails to explain the near-constant measurement variation of y for low true concentration level (x) Rocke and Lorenzato [1995]. To fix these two problems at the same time, Bhaumik and Gibbons [2005] proposed a log-normal model that combines both types of errors:

$$y_{ijk} = \alpha_i + \beta_i x_j e^{\eta_{ijk}} + e_{ijk}, \tag{2}$$

where y_{ijk} is the kth measurement at the jth concentration level in the i th laboratory, $i = 1, 2, ..., q, j = 1, 2, ..., r, k = 1, 2, ..., N_{ij}$, x_j is the true concentration at the jth level, and α_i and β_i are the calibration parameters for the ith laboratory.

In this model the η_{ijk}'s represent proportional error at higher true concentrations, and we assume that the distributions of the proportional errors remain the same for all laboratories.

Note that for this particular model, when x is 0 or near 0, the model is reduced to:

$$\begin{aligned} y_{ijk} &= \alpha_i + \beta_j \times 0 \times e^{\eta_{ijk}} + e_{ijk} \\ y_{ijk} &= \alpha_i + e_{ijk} \end{aligned} \tag{3}$$

4 Analysis

We choose the Method of Moments to estimate the model parameters. We choose the Method of Moments because it is straightforward and easy to implement when observations with lower concentrations are available, which happens to be the case with our Interlaboratory Data for Cadmium. Besides, the estimates obtained by Method of Moments are asymptotically efficient.

4.1 The Parameter Estimation Procedures (Using Data Separation Technique)

4.1.1 (Partial) Estimation Using Low-Concentration Observations

We estimate the variance σ_e^2 of the additive errors e_{ijk}'s (of the kth measurement at the jth concentration level in the i th laboratory, $i = 1, 2, ..., q, j = 1, 2, ..., r, k = 1, 2, ..., N_{ij}$) that are present primarily at low-level concentrations and the calibration parameter α_i for the ith laboratory using low-concentration observations.

Assuming that observations corresponding to zero or near-zero are available. Let y_{i0k} be the kth measured concentration corresponding to the true low-level concentration from laboratory i and n_{i0} is the number of samples with true low-level concentrations submitted to laboratory i:

(1) Estimate σ_e^2 by the variance of the observations with zero or near-zero concentrations (y_{i0k}) by

$$\hat{\sigma}_e^2 = \frac{1}{q} \sum_{i=1}^{q} \left(\frac{\sum_{k=1}^{n_{i0}} (y_{i0k} - \bar{y}_{i0})^2}{n_{i0} - 1} \right) \tag{4}$$

(2) Estimate α_i from the observations with zero or near-zero concentrations (y_{i0k}) by

$$\hat{\alpha}_i = \frac{\sum_{k=1}^{n_{i0}} y_{i0k}}{n_{i0}} \tag{5}$$

4.1.2 (Partial) Estimation Using Higher-Concentration Observations

(1) Let $z_{ijk} = \frac{(y_{ijk}-\alpha_i)}{x_j}$ and $u_{ijk} = \frac{e_{ijk}}{xj}$.

Using the estimate $\hat{\alpha}_i$ of α_i we compute the z_{ijk}'s by

$$\begin{aligned} y_{ijk} &= \alpha_i + \beta_i x_j e^{\eta_{ijk}} + e_{ijk} \\ y_{ijk} - \alpha_i &= \beta_i x_j e^{\eta_{ijk}} + e_{ijk} \\ \frac{y_{ijk} - \alpha_i}{x_j} &= \beta_i e^{\eta_{ijk}} + \frac{e_{ijk}}{x_j} \text{ (given } x_j's \text{ have higher concentrations)} \\ z_{ijk} &= \beta_i e^{\eta_{ijk}} + \mu_{ijk} \\ \frac{z_{ijk}}{\beta_i} &= e^{\eta_{ijk}} + \frac{\mu_{ijk}}{\beta_i} \text{ (given } \beta_i \neq 0). \end{aligned} \tag{6}$$

(2) Let

$$\gamma = E(e^{\eta}) = e^{\sigma_\eta^2/2},\ \mu_{zi} = E(z_{ijk}),\ \sigma_{zi}^2 = \frac{\sum_{j=1}^{r} var(z_{ijk})}{r},\ \sigma_\mu^2 = \frac{\sigma_e^2 \sum_{j=1}^{r} 1/x_j^2}{r} \tag{7}$$

Using equation (6) we can calculate

$$\frac{\mu_{zi}}{\beta_i} = \gamma \text{ and } \frac{\sigma_{zi}^2}{\beta_i^2} = (\gamma^4 - \gamma^2) + \frac{\sigma_\mu^2}{\beta_i^2}. \tag{8}$$

(3) Replacing γ by $\frac{\mu_{zi}}{\beta_i}$ in the second part of equation (8) we obtain the following equation:

$$\beta_i = \sqrt{\frac{\mu_{zi}^4}{\sigma_{zi}^2 - \sigma_\mu^2 + \mu_{zi}^2}} \text{ and } \sigma_\eta^2 = \frac{2\sum_{i=1}^{q} ln(\mu_{zi}/\beta_i)}{q}. \tag{9}$$

To estimate β_i and σ_η from equation (9), we need estimates of μ_{zi}, σ_{zi}^2, and σ_μ^2 (the error variance of the model (6)).

(4) Let n_{ij} be the number of observations with higher concentrations collected from the ith laboratory for concentration level j and $n_i = \sum_{j=1}^{r} \eta_{ij}$, then

$$\begin{aligned}\hat{\mu}_{z_{ij}} &= \frac{\sum_{k=1}^{n_{ij}} z_{ijk}}{n_{ij}},\ \hat{\mu}_{z_i} = \frac{\sum_{j=1}^{r} \hat{\mu}_{z_{ij}}}{r}, \\ \hat{\sigma}^2_{zi} &= \frac{\sum_{j=1}^{r}\sum_{k=1}^{\eta_{ij}}(z_{ijk} - \hat{\mu}_{z_{ij}})^2/(\eta_{ij}-1)}{r},\ \hat{\sigma}^2_{\mu} = \frac{\sum_{j=1}^{r} \hat{\sigma}^2_e/x_j^2}{r}\end{aligned} \tag{10}$$

(5) Therefore We can calculate the estimates of the variance of the proportional error at higher true concentrations σ^2_η and the calibration parameter β_i for the ith laboratory as

$$\hat{\beta}_i = \sqrt{\frac{\hat{\mu}^4_{z_i}}{\hat{\sigma}^2_{z_i} - \hat{\sigma}^2_{\mu} + \hat{\mu}^2_{z_i}}} \text{ and } \hat{\sigma}^2_{\eta} = \frac{2\sum_{i=1}^{q} ln(\hat{\mu}_{z_i}/\hat{\beta}_i)}{q} \tag{11}$$

4.2 Point Estimation Of X

```
## [1] 7.896

## [1] 0.8720 0.8682 1.1870 0.8877 0.9545

## [1] 0.9914 0.8865 1.0176 0.9121 0.9751

## [1] 0.9317 0.8774 1.1023 0.8999 0.9648

## [1] 0.0661450 0.0008652 0.1370575 0.0124709 0.0010388

## [1] 0.0037388 0.0004391 0.0155757 0.0013298 0.0025826

## [1] 0.9187 0.8829 1.0735 0.9018 0.9692

## [1]  0.6200  0.0700 -6.6800 -0.6308 -0.1924

## [1] 0.9187 0.8829 1.0735 0.9018 0.9692

##           [,1]       [,2]     [,3]      [,4]     [,5]       [,6]
## mu_zij20  "20 ug/L"  "0.872"  "0.8682"  "1.187"  "0.88774"  "0.95452"
## mu_zij100 "100 ug/L" "0.9914" "0.88652" "1.0176" "0.912118" "0.975092"

## [1] 0.9317 0.8774 1.1023 0.8999 0.9648

## [1] 0.0349419 0.0006522 0.0763166 0.0069003 0.0018107

## [1] 7.896

## [1] 0.01026

## [1] 0.01102

## [1] 1.006
```

4.3 Point Estimation of X

Suppose that the same analyte with an unknown true concentration X is tested in $q^{'}$ independent laboratories, and that $Y_1, Y_2, ..., Y_{q'}$ are the corresponding new observations. We want to compute a point estimate of X using the available information from each of the $q^{'}$ laboratories and then combine them. Following model (2), we can obtain

$$E(Y_i) = \alpha_i + \beta_i X \gamma,$$

and hence an estimate of X from the ith laboratory, denoted by $\hat{X}_i$, is

$$\hat{X}_i = \frac{Y_i - \hat{\alpha}_i}{\hat{\beta}_i \hat{\gamma}} \text{ and } \hat{X} = \sum_{i=1}^{q^{'}} \frac{\hat{X}_i}{q^{'}}, \tag{12}$$

where $\hat{X}$ is the combined estimate of X. $\hat{X}_i$ is asymptotically unbiased, and the asymptotic expression of the variance of $\hat{X}$ is

$$\begin{aligned} var(\hat{X}) &= E[var(\hat{X}/y)] + var[E[\hat{X}]/y] \\ &= \frac{\sum_{i=1}^{q^{'}} \sigma_e^2/(\beta_i^2\gamma^2)(1 + 1/n_{0i})}{q^{'2}} + \frac{X^2(\gamma^2 - 1)}{q^{'}}, \end{aligned} \tag{13}$$

where y is the vector of past observations, $var(\hat{X})/y$ is the conditional variance of $\hat{X}$ given y, and $E(\hat{X}/y)$ is the conditional expectation of $\hat{X}$ given y. The variance of $\hat{X}$ in equation (13) depends on X and increases with increasing levels of concentrations.

4.4 Variance and Confidence Interval of X

```
############################ estimation of x ################################
########## choose 1st observation of each lab as new observation ##########

## 0 ug / L

## point estimate

y1 <- low[1,]
y2 <- low[6,]
y3 <- low[11,]

x <- function(y, alpha, beta, gamma) {
  xi_hat <- (y - alpha) / (beta * gamma)
```

```
}

x1 <- x(y1$Real, y1$alpha_i, y1$beta_i, gamma)
x2 <- x(y2$Real, y2$alpha_i, y2$beta_i, gamma)
x3 <- x(y3$Real, y3$alpha_i, y3$beta_i, gamma)
x_hat <- (x1 + x2 + x3) / 3
x_hat
```

```
## [1] -1.577
```

```
## variance

variance <- (sigma_e2 / (y1$beta_i^2 * gamma^2)*(1 + 1 / 5) + sigma_e2 / (y2$beta_i^2 *
                sigma_e2 / (y2$beta_i^2 * gamma^2)*(1 + 1 / 5)) / 3^2 +
  0^2 * (gamma^2 - 1) / 3

## confidence interval
y_bar <- ( y1$Real + y2$Real + y3$Real ) / 3
sigma_alpha2 <- var(alpha_i)

cll <- max(0, y_bar - 1.96*sqrt((sigma_e2 + sigma_alpha2)/ 3))
clu <- y_bar + 1.96 * sqrt((sigma_e2 + sigma_alpha2) /3)

CL_low <- c(cll, clu)

# 20 ug / L

## point estimate

y12 <- twenty[1,]
y22 <- twenty[6,]
y32 <- twenty[11,]

x12 <- x(y12$Real, y12$alpha_i, y12$beta_i, gamma)
x22 <- x(y22$Real, y22$alpha_i, y22$beta_i, gamma)
x32 <- x(y32$Real, y32$alpha_i, y32$beta_i, gamma)
x_hat2 <- (x12 + x22 + x32) / 3
x_hat2
```

```
## [1] 20.48
```

```
## variance

variance2 <- (sigma_e2 / (y12$beta_i^2 * gamma^2)*(1 + 1 / 5) + sigma_e2 / (y22$beta_i^2
  sigma_e2 / (y32$beta_i^2 * gamma^2)*(1 + 1 / 5)) / 3^2 +
  20^2 * (gamma^2 - 1) / 3
```

```
## confidence interval

X1 <- 20

c1i <- beta_i^2 * X1^2 * (gamma^4 - gamma^2) + sigma_e2
c2i <- c1i / (beta_i^2*X1^2)
c3i <- log((1 + sqrt(1 + 4*c2i)) / 2)

sum1 <- log(y12$Real - alpha_i[1]) / sqrt(c3i[1]) +
  log(y22$Real - alpha_i[2]) / sqrt(c3i[2]) +
  log(y32$Real - alpha_i[3]) / sqrt(c3i[3])

# lower bound

f1 <- function (x) {
  sum1- log(beta_i[1] * x) / sqrt(c3i[1]) -
    log(beta_i[2] * x) / sqrt(c3i[2]) -
    log(beta_i[3] * x) / sqrt(c3i[3]) -
    1.96*sqrt(3)
}

# solve for x

cll20 <- uniroot(f1, lower = 0.1, upper = 10000000)$root

# upper bound

f2 <- function (x) {
  sum1- log(beta_i[1] * x) / sqrt(c3i[1]) -
    log(beta_i[2] * x) / sqrt(c3i[2]) -
    log(beta_i[3] * x) / sqrt(c3i[3]) +
    1.96*sqrt(3)
}

# solve for x

clu20<- uniroot(f2, lower = 0.1, upper = 10000000)$root

CL_20 <- c(cll20, clu20)

# 100 ug / L

## point estimate
```

```r
y13 <- hundred[1,]
y23 <- hundred[6,]
y33 <- hundred[11,]

x13 <- x(y13$Real, y13$alpha_i, y13$beta_i, gamma)
x23 <- x(y23$Real, y23$alpha_i, y23$beta_i, gamma)
x33 <- x(y33$Real, y33$alpha_i, y33$beta_i, gamma)
x_hat3 <- (x13 + x23 + x33) / 3
x_hat3
```

```
## [1] 102.1
```

```r
## variance

variance3 <- (sigma_e2 / (y13$beta_i^2 * gamma^2)*(1 + 1 / 5) + sigma_e2 / (y23$beta_i^2
                sigma_e2 / (y33$beta_i^2 * gamma^2)*(1 + 1 / 5)) / 3^2 +
  100^2 * (gamma^2 - 1) / 3

## confidence interval

X2 <- 100

c1i2 <- beta_i^2 * X2^2 * (gamma^4 - gamma^2) + sigma_e2
c2i2 <- c1i2 / (beta_i^2*X2^2)
c3i2 <- log((1 + sqrt(1 + 4*c2i2)) / 2)

sum2 <- log(y13$Real - alpha_i[1]) / sqrt(c3i2[1]) +
  log(y23$Real - alpha_i[2]) / sqrt(c3i2[2]) +
  log(y33$Real - alpha_i[3]) / sqrt(c3i2[3])

# lower bound

f3 <- function (x) {
  sum2- log(beta_i[1] * x) / sqrt(c3i2[1]) -
    log(beta_i[2] * x) / sqrt(c3i2[2]) -
    log(beta_i[3] * x) / sqrt(c3i2[3]) -
    1.96*sqrt(3)
}

# solve for x

cll100 <- uniroot(f3, lower = 0.1, upper = 10000000)$root

# upper bound
```

```r
f4 <- function (x) {
  sum2- log(beta_i[1] * x) / sqrt(c3i2[1]) -
    log(beta_i[2] * x) / sqrt(c3i2[2]) -
    log(beta_i[3] * x) / sqrt(c3i2[3]) +
    1.96*sqrt(3)
}

# solve for x

clu100 <- uniroot(f4, lower = 0.1, upper = 10000000)$root

CL_100 <- c(cll100, clu100)
```

Variance for X is 3.905, 4.9507, 40.4201 for 0, 20, 100 $\mu g/L$ true concentration level, respectively.

Bhaumik and Gibbons [2005] proposed a method for constructing confidence regions, which is an approximation based on a normal or lognormal distribution.

Firstly, to construct a confidence region for low-level concentrations, let Y_{i0} be an observation collected from the ith laboratory with low-level true concentration. Define

$$\bar{Y_0} = \sum_{i=1}^{q'} Y_{i0}/n_{0\cdot},$$

where n_0 is the total number of measurements for low-level true concentration from all q' laboratories. For a low-level true concentration X_0, the $(1-\alpha)100\%$ confidence region of X_0 is

$$(max(0, \bar{Y_0} - z_{\alpha/2}\sqrt{(\hat{\sigma}_e^2 + \hat{\sigma}_\alpha^2)/n_0}), \bar{y}_0 + z_{\alpha/2}\sqrt{(\hat{\sigma}_e^2 + \hat{\sigma}_\alpha^2)/n_0}).$$

$\hat{\sigma}_\alpha^2$ represents the variability of α_i across all q laboratories in the calibration sample.

Secondly, to construct a confidence region for larger X, Bhaumik and Gibbons [2005] used the following lognormal approximation. Let

$$c_{1i} = var(Y_i) = \beta_i^2 X^2(\gamma^4 - \gamma^2) + \sigma_e^2, c_{2i} = \frac{c_{1i}}{\beta_2^i X^2}, and\ c_{3i} = ln(\frac{1 + \sqrt{1 + 4c_{2i}}}{2}).$$

Bhaumik and Gibbons [2005] also showed in a lemma that the approximate variance of $ln(\frac{Y-\alpha_i}{\beta_i X})$ is c_{3i}.

Lemma 1. Suppose that for an unknown concentration X, the corresponding observation Y_i collected from the ith laboratory follows model (2). Define $V_i = ln(\frac{Y_i-\alpha_i}{\beta_i X})$. For a larger concentration X, the approximate variance of V_i is c_{3i}.

Proof. For a larger concentration X, the corresponding e_i in model (2) becomes insignificant compared to Xe^{η_i}, and hence the approximate distribution of $(\frac{Y_i-\alpha_i}{\beta_i X})$ is lognormal. Using the expression for the variance of a lognormal distribution, $c_{2i} = var(\frac{Y_i-\alpha_i}{\beta_i X})$ can be expressed as $v^2 - v$, for a positive number v. The positive root of the quadratic equation $c_{2i} = v^2 - v$ is $\frac{1+\sqrt{1+4c_{2i}}}{2}$, and hence the variance of $ln(\frac{Y_i-\alpha_i}{\beta_i X})$ is c_{3i}.

Let

$$Z_i(X) = \frac{ln((Y_i - \alpha_i)/\beta_i X)}{\sqrt{c_{3i}}} = \frac{ln(Y_i - \alpha_i) - ln(\beta_i X)}{\sqrt{c_{3i}}}. \tag{14}$$

Thus $Z_i(X) \sim N(0,1)$ and the approximate distribution of $Z(X) = \sum_{i=1}^{q'} Z_i(X)/\sqrt{q'} \sim N(0,1)$, where $N(0,1)$ denotes a standard normal distribution. We replace the parameters on the right side of (14) by their corresponding estimates to compute their numerical values. Thus the $(1-\alpha)100\%$ confidence region for X is

$$\mathcal{R}(X) = \left\{X : -Z_{\alpha/2} \leq Z(X) \leq Z_{\alpha/2}\right\}. \tag{15}$$

5 Results and Discussions

```
## make a table
x <- matrix(round(c(x_hat, variance, cll, clu, scl_low, x_hat2,
                    variance2, cll20, clu20, NA,
                    x_hat3, variance3, cll100, clu100, NA), 4),
            3, 5, byrow = T)
x <- as.matrix(cbind(c('0 ug/L', '20 ug/L', '100 ug/L'), x))
colnames(x) <- c('True Concentration','X_hat', 'Var(X)', 'Lower Bound',
                 'Upper Bound', 'SCL')
kable(x, format = 'pandoc')
```

True Concentration	X_hat	Var(X)	Lower Bound	Upper Bound	SCL
0 ug/L	-1.5773	3.905	0	1.1703	0.953
20 ug/L	20.4786	4.9507	15.4935	23.1297	NA
100 ug/L	102.1374	40.4201	90.7669	116.1489	NA

References

Dulal K Bhaumik and Robert D Gibbons. Confidence regions for random-effects calibration curves with heteroscedastic errors. *Technometrics*, 47(2):223–231, may 2005.

David M. Rocke and Stefan Lorenzato. A two-component model for measurement error in analytical chemistry. *Technometrics*, 37(2):176–184, may 1995.