BIOSTAT 571 Homework 1

Wenxiao Gu

February 5, 2014

Contents

1 Problem 1

Consider simple linear regression with a single covariate and an intercept, such that we are fitting

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

with scalar observations $Y_1, \ldots, Y_n$. It is known that the OLS estimator of β_1 can be written as a weighted average of the pairwise slopes $(Y_i - Y_j)/(x_i - x_j)$. This is called the "average slope" representation.

In last homework, the Gauss-Markov gives us the unbiased estimator with minumum variance if W is symmatric and positive definite. In Wakefield's notation, W is the same as V.

In this homework, our working covariance matrix V is set to be diagonal with known entries $\sigma_1^2, \ldots, \sigma_n^2$. Thus, we have the optimal optimal GLS estimate of β_1.

Let $W = V^{-1}$.

$$\begin{aligned}
X^T V^{-1} X &= X^T \begin{pmatrix} \sigma_1^{-2} & & \\ & \ddots & \\ & & \sigma_n^{-2} \end{pmatrix} X \\
&= \begin{bmatrix} 1, \ldots, 1 \\ x_1, \ldots, x_n \end{bmatrix} \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix} \begin{bmatrix} 1, \ldots, 1 \\ x_1, \ldots, x_n \end{bmatrix}^T \\
&= \begin{bmatrix} \sum_i (w_i) & \sum_i (w_i x_i) \\ \sum_i (w_i x_i) & \sum_i (w_i x_i^2) \end{bmatrix}
\end{aligned}$$

$$\left(X^T V^{-1} X\right)^{-1} = \frac{1}{\left\{ \left[\sum_i (w_i)\right] \left[\sum_i \left(w_i x_i^2\right)\right] - \left[\sum_i (w_i x_i)\right]^2 \right\}} \begin{bmatrix} \sum_i \left(w_i x_i^2\right) & -\sum_i (w_i x_i) \\ -\sum_i (w_i x_i) & \sum_i (w_i) \end{bmatrix}$$

$$\begin{aligned}
X^T V^{-1} Y &= \sum_j \left[w_j x_j y_j\right] \\
&= \begin{bmatrix} 1, \ldots, 1 \\ x_1, \ldots, x_n \end{bmatrix} \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix} \begin{bmatrix} y_1, \ldots, y_n \end{bmatrix}^T \\
&= \begin{bmatrix} \sum_i (w_i y_i) \\ \sum_i (w_i x_i y_i) \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta} &= \left(X^T V^{-1} X\right)^{-1} \left(X^T V^{-1} Y\right) \\
&= \frac{1}{\left[\sum_i (w_i)\right] \left[\sum_i \left(w_i x_i^2\right)\right] - \left[\sum_i (w_i x_i)\right]^2} \begin{bmatrix} \sum_i \left(w_i x_i^2\right) & -\sum_i (w_i x_i) \\ -\sum_i (w_i x_i) & \sum_i (w_i) \end{bmatrix} \begin{bmatrix} \sum_i (w_i y_i) \\ \sum_i (w_i x_i y_i) \end{bmatrix} \\
\hat{\beta}_1 &= \frac{\left[-\sum_i (w_i x_i)\right] \left[\sum_i (w_i y_i)\right] + \left[\sum_i (w_i)\right] \left[\sum_i (w_i x_i y_i)\right]}{\left[\sum_i (w_i)\right] \left[\sum_i \left(w_i x_i^2\right)\right] - \left[\sum_i (w_i x_i)\right]^2} = \frac{A}{D}
\end{aligned}$$

Next do some trick from STAT 570 HW2. The numerator and the denominator have similar forms of x and y. They should be calculated from different orders of summation.

$$
\begin{aligned}
A &= \left[-\sum_i (w_i x_i)\right]\left[\sum_i (w_i y_i)\right] + \left[\sum_i (w_i)\right]\left[\sum_i (w_i x_i y_i)\right] \\
&= -\sum_j (w_j x_j) \sum_i (w_i y_i) + \sum_j (w_j) \sum_i (w_i x_i y_i) \\
&= -\sum_i \sum_j w_j x_j w_i y_i + \sum_i \sum_j w_i w_j x_i y_i \\
&= \sum_i \sum_j (w_i w_j x_i y_i - w_i w_j x_j y_i) \\
&= \frac{1}{2}\sum_i \sum_j (w_i w_j x_i y_i - w_i w_j x_j y_i) + \frac{1}{2}\sum_i \sum_j (w_i w_j x_j y_j - w_i w_j x_i y_j) \\
&= \frac{1}{2}\sum_i \sum_j w_i w_j (x_i y_i - x_j y_i + x_j y_j - x_i y_j) \\
&= \frac{1}{2}\sum_i \sum_j w_i w_j (x_i - x_j)(y_i - y_j) \text{ trick from STAT570 HW2 Solution} \\
&= \frac{1}{2}\sum_i \sum_j w_i w_j (x_i - x_j)^2 \frac{(y_i - y_j)}{(x_i - x_j)} \\
D &= \left[\sum_i (w_i)\right]\left[\sum_i \left(w_i x_i^2\right)\right] - \left[\sum_i (w_i x_i)\right]^2 \\
&= \frac{1}{2}\sum_i \sum_j w_i w_j (x_i - x_j)^2 \frac{(x_i - x_j)}{(x_i - x_j)} \\
&= \frac{1}{2}\sum_i \sum_j w_i w_j (x_i - x_j)^2
\end{aligned}
$$

Hence the estimate would be as following,

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i \sum_j w_i w_j (x_i - x_j)(y_i - y_j)}{\sum_i \sum_j w_i w_j (x_i - x_j)^2} \\
&= \sum_i \sum_j \frac{w_i w_j (x_i - x_j)^2}{\sum_i \sum_j w_i w_j (x_i - x_j)^2} \cdot \frac{(y_i - y_j)}{(x_i - x_j)} \\
&= \sum_i \sum_j \frac{w_i w_j (x_i - x_j)^2}{\sum_i \sum_j w_i w_j (x_i - x_j)^2} \cdot Slope[i, j]
\end{aligned}
$$

The optimal GLS estimate of β_1 is a weighted average of the pairwise slopes at the interval i and j. The variance for each point in a pair and the distance square between two points are both proportional to the weight, and the weight for the pair is higher if the varaince for each point is smaller.

$$
\frac{w_i w_j (x_i - x_j)^2}{\sum_i \sum_j w_i w_j (x_i - x_j)^2}.
$$

The slope is $\frac{(y_i - y_j)}{(x_i - x_j)}$. From the expression, we see an extreme case. If the two points are both at $x = x$, then no information in this slope would lead to the estimate of β_1.

For the case with a non-independent covariance matrix V, the weight for each pair would be very similar to the independent case. The simulation in R is in the Appendix.

$$\begin{aligned} Var(Slope[i,j]) &= \frac{Var(y_i - y_j)}{(x_i - x_j)^2} \\ &= \frac{\sigma_i^2 + \sigma_j^2 - 2cov(y_j, y_i)}{(x_i - x_j)^2} \\ &= \frac{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j}{(x_i - x_j)^2} \end{aligned}$$

We can see that the variance of the slope for each pair would be larger if the points are positively correlated, and would be smaller if negatively correlated.

2 Problem 2

2.1 (a)

$$Y_{ij} = \gamma_0 + \gamma_1 x_i + \gamma_2 z_{ij} + \gamma_3 x_i z_{ij} + \epsilon_{ij} \tag{1}$$

Let $i = 1, ..., n$ be the index of subjects,$j = 1, ..., m$ be the index of measurements for each i.
$E[\gamma_{ij}|z_{ij}] = 0$. Thus

$$\begin{aligned} E[Y_{ij}|x_i,\ z_{ij}] &= E[\hat{\alpha}_{0i} + \hat{\alpha}_{1i} z_{ij} + \epsilon_{ij}|x_i,\ z_{ij}] \\ &= \alpha_{0i} + z_{ij} E[\hat{\alpha}_{1i}|x_i] \\ &= \alpha_{0i} + z_{ij} E[\hat{\beta}_0 + \hat{\beta}_1 x_i + error_i|x_i] \\ &= \alpha_{0i} + z_{ij}(\beta_0 + \beta_1 x_i) \\ &= \alpha_{0i} + \beta_0 z_{ij} + \beta_0 z_{ij} x_i \end{aligned}$$

Hence $E(\hat{\beta}_1) = \gamma_3$. So $\hat{\beta}_1$ is the consistent estimator of γ_3.

2.2 (b)

	Balanced, Complete $\alpha = 0$	Balanced, Complete $\alpha \neq 0$	Balanced, Incomplete $\alpha = 0$	Balanced, Incomplete $\alpha \neq 0$
One-stage OLS (M-B se)	Valid (3.)	Invalid (1.)	Valid (3.)	Invalid (1.)
One-stage OLS (Sand se)	Valid (3.)	Invalid (1.)	Valid (3.)	Invalid (1.)
Two-stage OLS (M-B se)	Valid (1.)	Valid (1.)	Invalid (2.)	Invalid (2.)
Two-stage OLS (Sand se)	Valid(1.)	Valid (1.)	Valid (2.)	Valid (2.)

In the above cases, all the mean models are correct.

1. Sandwich estimates always give asymptotically valid CIs if the data are uncorrelated, and if the the mean model is correct.
   - In two-stage OLS, the errors are always uncorrelated. They will produce the vallid standard error.
   - The correlated ($\alpha \neq 0$) in one-stage OLS model will not produce valid standard error.
2. In the two-stage OLS, if the data are balanced and complete, we got the same precision for $\hat{\alpha}$. As we mentioned earlier, the errors in the second stage are uncorrelated. Therefore, all the assumptions for model-based estimates are satisfied, and we got valid model based CI estimates. If the data are not complete, we got dierent precisions for dierent $\hat{\alpha}$. The error in the second stage is thus heteroscedastic. So the model based CI estimates produce invalid standard errors.
3. In the one-stage OLS, the errors are uncorrelated if and only if $\alpha = 0$. So they will produce the vallid standard error if and only if $\alpha = 0$.

2.3 (c)

Neither of the following two methods is appropriate for estimation of parameters associated with covariates that vary within clusters.

- the overdispersed binomial method for correlated Bernoulli data
  - In traditional binomial methods, we assume that the mean-variance relationship is following the one of binomial distribution. In the over-dispersed binomial method, we aggregate the bernoulli data for individuals into binomial counts, define a dispersion parameter, and allow the variance to be proportional to the mean. This method resolves the problem of correlated data if the mean model and the proportionality assumptions are correct. However, it assumes the same overdispersion factor for all clusters/groups and cannot adjust different sizes of groups.
  - For the over-dispersed binomial model, the model based standard error is valid in each of the designs.
- The two-stage model allows us to estimate a summary parameter for each group/cluster, and then estimate the effect of that parameter across across clusters with different values.
  - When the design is balanced and complete, and with homoscedasticity for each cluster, the model based standard error in the second stage of the 2-stage OLS will be valid, because we the mean model and the mean-variance relationship are both correct. However, the model-based standard error estimates in the first stage are still invalid because we have correlated observations.
  - If the design is balanced and incomplete, then model-based standard error estimates is no longer valid because in the first stage the data are correlated, and in the second stage there is heteroscedasticity.

3 Problem 3

3.1 (a)

From the notes, we can see the covariance matrix of an OLS estimmate is,

$$Cov(\hat{\beta}_{OLS}) = [\sum_{i=1}^{n}(X_i^T X_i)]^{-1} \cdot [\sum_{i=1}^{n}(X_i^T \Sigma_i X_i)] \cdot [\sum_{i=1}^{n}(X_i^T X_i)]^{-1}$$

and the covariance of the optimal GLS estimate is

$$Cov(\hat{\beta}_{GLS}) \quad = \quad [\sum_{i=1}^{n}(X_i^T \Sigma_i^{-1} X_i)]^{-1}$$

So, the relative efficiency for estimating β_0 is $e(\beta_0) = \frac{Cov(\hat{\beta}_{GLS})_{[1,1]}}{Cov(\hat{\beta}_{OLS})_{[1,1]}}$, while the relative efficiency for estimating β_1 is $e(\beta_1) = \frac{Cov(\hat{\beta}_{GLS})_{[2,2]}}{Cov(\hat{\beta}_{OLS})_{[2,2]}}$.

In the our example of data, we have $n = 5$, and if $X_i \equiv X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ x_1 & x_2 & x_3 & x_4 & x_5 \end{pmatrix}$ and the covariance matrix as following,

$$
\begin{aligned}
\Sigma_i &\equiv \Sigma \\
&= \sigma^2 \begin{bmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \rho^{|t_1-t_4|} & \rho^{|t_1-t_5|} \\ \rho^{|t_1-t_2|} & 1 & \rho^{|t_2-t_3|} & \rho^{|t_2-t_4|} & \rho^{|t_2-t_5|} \\ \rho^{|t_1-t_3|} & \rho^{|t_2-t_3|} & 1 & \rho^{|t_3-t_4|} & \rho^{|t_3-t_5|} \\ \rho^{|t_1-t_4|} & \rho^{|t_2-t_4|} & \rho^{|t_3-t_4|} & 1 & \rho^{|t_4-t_5|} \\ \rho^{|t_1-t_5|} & \rho^{|t_2-t_5|} & \rho^{|t_3-t_5|} & \rho^{|t_4-t_5|} & 1 \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 & \rho^4 \\ \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 \\ \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 \\ \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 \\ \rho^4 & \rho^3 & \rho^2 & \rho^1 & 1 \end{bmatrix}
\end{aligned}
$$

Thus the data can be put into the following formulae,

$$
\begin{cases} Cov(\hat{\beta}_{OLS}) = [\sum_{i=1}^n (X_i^T X_i)]^{-1} \cdot [\sum_{i=1}^n (X_i^T \Sigma_i X_i)] \cdot [\sum_{i=1}^n (X_i^T X_i)]^{-1} \\ Cov(\hat{\beta}_{GLS}) = [\sum_{i=1}^n (X_i^T \Sigma_i^{-1} X_i)]^{-1}. \end{cases}
$$

Hence the relative efficiency for estimating β_0 and β_1 is more accurate sssasasa

$$
\frac{Cov(\hat{\beta}_{GLS})}{Cov(\hat{\beta}_{OLS})} = \frac{[\sum_{i=1}^n (X_i^T \Sigma_i^{-1} X_i)]^{-1}}{[\sum_{i=1}^n (X_i^T X_i)]^{-1} \times [\sum_{i=1}^n (X_i^T \Sigma_i X_i)] \times [\sum_{i=1}^n (X_i^T X_i)]^{-1}}
$$

3.2 (b)

The following is the relative eciency of β_1 from OLS compared to GLS when $n = 10$, in each case of x.

$\rho =$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99	
x = (-2, -1, 0, 1, 2)	0.997	0.989	0.980	0.970	0.962	0.955	0.952	0.952	0.955	0.961	
x = (-1, -2, 0, 2, 1)	0.996	0.982	0.955	0.915	0.862	0.797	0.725	0.649	0.573	0.507	decrease
x = (0, -1, 1, 3, 2)	0.996	0.982	0.955	0.915	0.862	0.797	0.725	0.649	0.573	0.507	decrease
x = (0, -1, 1, 5, 2)	0.991	0.964	0.921	0.863	0.794	0.719	0.643	0.569	0.499	0.442	decrease

3.3 (c)

- Smaller values (<1) mean that the GLS estimate is better than OLS.
- GLS will produce more accurate values when the data are more correlated ($\rho > 0$). If the correlation is weak, the OLS will produce a relatively good result as the optimal GLS.
- In addition to the correlation, if the observations are far away from each other in x, then they would produce a more accurate pairwise slope, meaning lower variance of the estimated slope. Thus the OLS would produce a large weight for this particular slope. However, the GLS will consider the correlation as well. In our case, the correlation is small if the two points are far away from each other, meaning that in this case the GLS may not be as good as OLS. Hence the OLS will be relatively good compared to the optimal GLS if the distance contains less information than the correlation.

Now let's look at the table. The following are two examples of calculation.
In the first row of relative efficiency, the maximum distance is 2-(-2) = 4, and their correlation is ρ^4.
In the 2nd row of relative efficiency, the maximum distance is 2-(-2) = 4, and their correlation is ρ^2.

1. **Why are the relative efficiencies in the first row similar?** Because their distance is large and the correlation is small, the OLS gives as good as result as the GLS gives as we talked above.

2. **Why are the relative efficiencies in the 2nd row and in 3rd row similar with respect to each correlation parameter ρ?** Because in the 2nd row the correlation is larger and the OLS have no information from the correlated data, thus the GLS gives better estimate.

3. **Why are the relative efficiencies in the 2nd row (in 3rd row, or in 4th row) increasing as ρ increases?** This is same to the 2nd question. The correlation gets larger and thus the GLS produces better estimate.

4. **Why are the relative efficiencies in the 4th row smaller than those in 3rd row with respect to each ρ?** Because the measurements at x_4 in the two rows are different, so the distance between x_4 and other points are longer than in 3rd row, while the pairwise correlations keep the same. As we talked above, the GLS would not produce good estimate if a pair of points with a long distance with a large correlation occurs.

Appendix

```
## pr1
m = 3
sig = seq(0.5, 1, l = m)
w = sig^(-2)
x = rnorm(m, 0, w)
x

beta0 = 0.5
beta1 = 1.2
y = beta0 + beta1 * x

X = cbind(rep(1, m), x)
Y = matrix(y, ncol = 1)
V = diag(w)
dim(w)
dim(X)

bhat = solve(t(X) %*% V %*% X) %*% (t(X) %*% V %*% Y)
bhat
slope = elem = matrix(0, m, m)

for (i in 1:m) {

    for (j in 1:m) {

        slope[i, j] = (y[i] - y[j])/(x[i] - x[j])
        elem[i, j] = (x[i] - x[j])^2 * (w[i] * w[j])
    }
}

sum(slope * elem, na.rm = 1)/sum(elem, na.rm = 1)

################################
m = 10
sig = seq(2, 2.5, l = m)
sig
sigma = matrix(-0.52, m, m)
diag(sigma) = sig^2
sigma[2, 1] = sigma[1, 2] = -1.95
sigma[(m - 1), m] = sigma[m, (m - 1)] = -1.95
sigma[m, 1] = sigma[1, m] = -1.95

sigma
library(mvtnorm)
x <- c(rmvnorm(n = 1, mean = rep(0, m), sigma = sigma))
x
y = beta0 + beta1 * x
X = cbind(rep(1, m), x)
Y = matrix(y, ncol = 1)
W = solve(sigma)
bhat = solve(t(X) %*% W %*% X) %*% (t(X) %*% W %*% Y)
bhat
```

```
slope = var.slop = elem = matrix(0, m, m)

for (i in 1:m) {
    for (j in 1:m) {
        var.slop[i, j] = (sigma[i, i] + sigma[j, j] - 2 * sigma[i, j])/(x[i] -
            x[j])^2
        slope[i, j] = (y[i] - y[j])/(x[i] - x[j])
        elem[i, j] = (x[i] - x[j])^2 * (w[i] * w[j])
    }
}
sum(slope * elem, na.rm = 1)/sum(elem, na.rm = 1)


## pr3
x = c(-2, -1, 0, 1, 2)
x = c(-1, -2, 0, 2, 1)
x = c(0, -1, 1, 3, 2)
x = c(0, -1, 1, 5, 2)

X = matrix(c(rep(1, 5), x), byrow = F, ncol = 2)
n = 10
ro <- c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99)
corr = matrix(0, ncol = 5, nrow = 5)
relative.effi = rep(0, n)
for (k in 1:10) {
    rho = ro[k]
    for (i in 1:5) {
        for (j in 1:5) {
            corr[i, j] = rho^(abs(i - j))
        }
    }
    covbeta_OLS = solve(t(X) %*% X) %*% (t(X) %*% corr %*% X) %*% solve(t(X) %*%
        X)/n
    covbeta_GLS = solve(t(X) %*% solve(corr) %*% X)/n
    relative.efficiency_1 = covbeta_GLS[2, 2]/covbeta_OLS[2, 2]
    relative.effi[k] = relative.efficiency_1
}

formatC(relative.effi, digit = 3)
```