BIOS 767 HW1

Mingwei Fei

January 18, 2023

1 Problem 1 - Computing moments

Consider the following model for subject i in group h at time j:

$$Y_{hij} = \mu + \alpha_h + \beta_j + \gamma_{hj} + b_{hi} + \epsilon_{hij}$$

where $b_{hi} \sim N(0, \sigma_b^2)$ and $\epsilon_{hij} \sim N(0, \sigma_e^2)$, for $h = 1, ..H, i = 1, ..n_h$, and $j = 1, ..J$. Assume that random variables that do not share a common value of h or i are independent.

(a) Compute $E[Y_{hij}]$.

$$\begin{aligned} Y_{hij} &= \mu + \alpha_h + \beta_j + \gamma_{hj} + b_{hi} + \epsilon_{hij} \\ E[Y_{hij}] &= \mu + \alpha_h + \beta_j + \gamma_{hj} + E[b_{hi}] + E[\epsilon_{hij}] \\ E[b_{hi}] &= 0, \qquad E[\epsilon_{hij}] = 0, \qquad \text{as} \quad b_{hi} \sim N(0, \sigma_b^2), \epsilon_{hij} \sim N(0, \sigma_e^2) \\ E[Y_{hij}] &= \mu + \alpha_h + \beta_j + \gamma_{hj} \end{aligned}$$

(b) Don't assume b_{hi} and ϵ_{hij} are independent. Derive an expression for $Var[Y_{hij}]$ that it as simplified as possible.

If b_{hi} and ϵ_{hij} are not independent,

$$\begin{aligned} Var[b_{hi} + \epsilon_{hij}] &= Var[b_{hi}] + Var[\epsilon_{hij}] + 2Cov[b_{hi}, \epsilon_{hij}] \\ Y_{hij} &= \mu + \alpha_h + \beta_j + \gamma_{hj} + b_{hi} + \epsilon_{hij} \\ Var[Y_{hij}] &= Var[b_{hi} + \epsilon_{hij}] \\ Var[b_{hi}] &= \sigma_b^2, \qquad Var[\epsilon_{hij}] = \sigma_e^2, \qquad \text{as} \quad b_{hi} \sim N(0, \sigma_b^2), \epsilon_{hij} \sim N(0, \sigma_e^2) \\ Var[Y_{hij}] &= \sigma_b^2 + \sigma_e^2 + 2Cov[b_{hi}, \epsilon_{hij}] \end{aligned}$$

(c) assume b_{hi} and ϵ_{hij} are independent. Derive an expression for $Var[Y_{hij}]$.

If b_{hi} and ϵ_{hij} are independent,

$$\begin{aligned} Cov[b_{hi}, \epsilon_{hij}] &= 0 \\ Var[Y_{hij}] &= \sigma_b^2 + \sigma_e^2 \end{aligned}$$

2 Problem 2 - Correlation in Data

Consider the general linear regression model:

$$Y_{n\times 1} = X_{n\times p}\beta + \epsilon_{n\times 1}$$

where ϵ is normal, and θ is unknown.

(a) Show the the ordinary least squares (OLS) estimator $\hat{\beta}^{OLS}$ is unbiased. We need to show $E(\hat{\beta}^{OLS}) = \beta$

$$\begin{aligned} \hat{\beta}^{OLS} &= (X^TX)^{-1}X^TY \\ E(\hat{\beta}^{OLS}) &= (X^TX)^{-1}X^TE(Y) = (X^TX)^{-1}X^TX\beta = \beta \end{aligned}$$

(b) Derive the variance of $\hat{\beta}^{OLS}$

$$\begin{aligned} Var(\hat{\beta}^{OLS}) &= Var((X^TX)^{-1}X^TY) = (X^TX)^{-1}X^TVar(Y)[(X^TX)^{-1}X^T]^T \\ Var(Y) &= \Sigma \\ Var(\hat{\beta}^{OLS}) &= (X^TX)^{-1}X^T\Sigma X[(X^TX)^{-1}] \end{aligned}$$

if X is non-singular, we can further simplify

$$\begin{aligned} Var(\hat{\beta}^{OLS}) &= X^{-1}(X^T)^{-1}X^T\Sigma XX^{-1}(X^T)^{-1} \\ &= X^{-1}\Sigma(X^T)^{-1} = [X^T\Sigma^{-1}X]^{-1} \end{aligned}$$

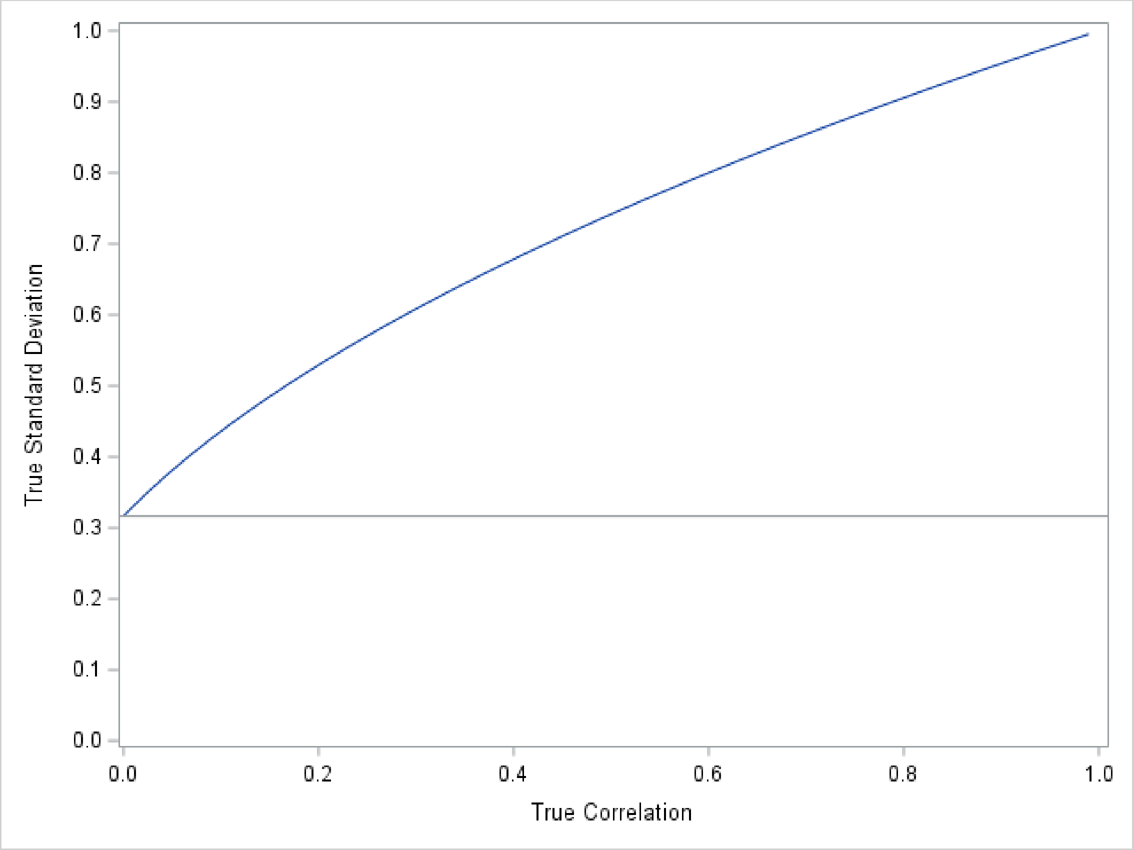
(c) Derive an expression for the variance of $\hat{\beta}^{OLS}$ (simplified as much as possible).

$$\begin{aligned} \Sigma &= \sigma^2 I_{n\times n} \\ Var(\hat{\beta}^{OLS}) &= [X^T\Sigma^{-1}X]^{-1} = \sigma^2(X^TX)^{-1} \end{aligned}$$

(d) Now, assume n = 10, p = 1, and let X be a vector of ones (e.g., an intercept only model). $\rho in(0,1)$, plot the true standard deviation of $\hat{\beta}^{OLS}$ as a function of ρ.

$$\Sigma = \sigma^2 I_{n\times n} = \begin{pmatrix} 1 & \rho & .. & \rho \\ \rho & 1 & .. & \rho \\ .. & .. & .. & \\ \rho & \rho & .. & 1 \end{pmatrix}$$

The plot of SD as a function of ρ as below:



(e) From the previous example, as $\rho \to 1$, you should observe that the true SD for $\hat{\beta}_{OLS} \to 1$. Concisely state why this is the case using one or two complete sentences.

From the SD formula, we calculate

$$\Sigma = \begin{pmatrix} 1 & \rho & .. & \rho \\ \rho & 1 & .. & \rho \\ .. & .. & .. & \\ \rho & \rho.. & . & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & .. & 1 \\ 1 & 1 & .. & 1 \\ .. & .. & .. & \\ 1 & 1 & .. & 1 \end{pmatrix}, \qquad \rho \to 1$$

$$\begin{aligned} X &= (1, ...1)^T, X^T X = 10 \\ Var(\hat{\beta}^{OLS}) &= (X^T X)^{-1} X^T \Sigma X [(X^T X)^{-1}] \\ &= 10^{-1}(1, ...1) \begin{pmatrix} 1 & 1 & .. & 1 \\ 1 & 1 & .. & 1 \\ .. & .. & .. & \\ 1 & 1 & .. & 1 \end{pmatrix}_{10\times 10} (1, ...1)^T 10^{-1} \\ &= 10^{-1}(10, ...10)(1, ...1)^T 10^{-1} = 1 \end{aligned}$$

When correlation coefficient is 1, we have the covariance is the same as variance, which shows that there is only one estimate, ie. overall mean. So the variance of the estimate is the variance of response variable or error term, which is 1.

(f) Hypothesis test of β. They do so using the OLS estimator but erroneously assume independent covariance matrix. when in fact the correlation between observations$\rho = \frac{1}{n-1}$. Note that, in this case is equal to the sample mean. What is the actual type I error rate for this test? What can you say about the impact of ignoring even small positive correlation in ananalysis when the sample size is large?

Assume we have the $\hat{\beta}$ based on the correlated covariance matrix. The incorrect variance is $\frac{1}{n}$ The p-value for two sided rejection region (type I error) is

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T \Sigma X [(X^T X)^{-1}] \\ p(x > \hat{\beta} | \beta = 0) &= p(x < -\hat{\beta} | \beta = 0) \\ Z &= \frac{\hat{\beta} - 0}{\sqrt{1/n}} \\ p &= 2p(x < -\hat{\beta} | \beta = 0) = 2CDF(-Z) \end{aligned}$$

So we will have inflated type I error if using the independent covariance matrix. The actual type I error rate is 0.16618 from the simulated code.

When sample size is large, the impact of ignoring the small positive correlation will become signficant as the true standard deviation is much larger than sd assuming independence $\sqrt{1/n}$. In turn, it would cause more inflated type I error.

3 Problem 3- Naive approach to handle correlation in Data

Consider the subject-specific linear regression model:

$$\begin{aligned} Y_{hij} &= \beta_{h0} + t_{hij}\beta_{h1} + \epsilon_{hij} \\ \hat{\beta}_{hi} &= (\hat{\beta}_{hi,0}, \hat{\beta}_{hi,1})^T = (X'_{hi} X_{hi})^{-1} X'_{hi} Y_{hi} \end{aligned}$$

The analyst plans to test the hypotheses

$$H_0 := \beta_{11} = \beta_{21} vs. H_1 : \beta_{11} \neq \beta_{21}$$

using a two-sample t-test using the values $\hat{\beta}_{hi,1} : h = 1, 2; i = 1, ..n$

(a) Compute $E[\hat{\beta}_{hi}], Var[\hat{\beta}_{hi}], Var[\hat{\beta}_{hi,1}]$.

$$\begin{aligned} \hat{\beta}_{hi} &= (X'_{hi} X_{hi})^{-1} X'_{hi} Y_{hi} \\ E(\hat{\beta}_{hi}) &= (X'_{hi} X_{hi})^{-1} X'_{hi} E[Y_{hi}] \\ &= (X'_{hi} X_{hi})^{-1} X'_{hi} X_{hi} \beta_{hi} = \beta_{hi} \end{aligned}$$

the variance

$$\begin{aligned}
Var(\hat{\beta}_{hi}) &= Var\Big((X'_{hi}X_{hi})^{-1}X'_{hi}Y_{hi}\Big) \\
&= (X'_{hi}X_{hi})^{-1}X'_{hi}Var(Y_{hi})\Big[(X'_{hi}X_{hi})^{-1}X'_{hi}\Big]^T \\
Var(Y) &= Var(\epsilon_{hi}) = \Sigma_{J_{hi}\times J_{hi}} \\
Var(\hat{\beta}_{hi}) &= (X'_{hi}X_{hi})^{-1}X'_{hi}\Sigma_{J_{hi}\times J_{hi}}\Big[(X'_{hi}X_{hi})^{-1}X'_{hi}\Big]^T \\
\hat{\beta}_{hi,1} &= (0,1)^T\hat{\beta}_{hi} \\
Var(\hat{\beta}_{hi,1}) &= Var\Big((0,1)^T\hat{\beta}_{hi}\Big) = (0,1)Var(\hat{\beta}_{hi})\begin{pmatrix}0\\1\end{pmatrix} \\
&= (0,1)(X'_{hi}X_{hi})^{-1}X'_{hi}\Sigma_{J_{hi}\times J_{hi}}\Big[(X'_{hi}X_{hi})^{-1}X'_{hi}\Big]^T\begin{pmatrix}0\\1\end{pmatrix}
\end{aligned}$$

(ii) Will the analyst's assumption of a common variance across the $\hat{\beta_{ji,1}}$ hold in general? It not, explain why not and give a sufficient condition such that the planned two-sample t-test's assumptions will hold.

The assumption of a common variance does not hold as it treated the different measurement for each subject follow the same distribution, while in reality there are heterogeinty of the measurement for each individual.

The sufficient condition for using t-test is to have the measurements follow identical distribution.