BIOS 767 HW1

Mingwei Fei

22 gennaio 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,

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

(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).

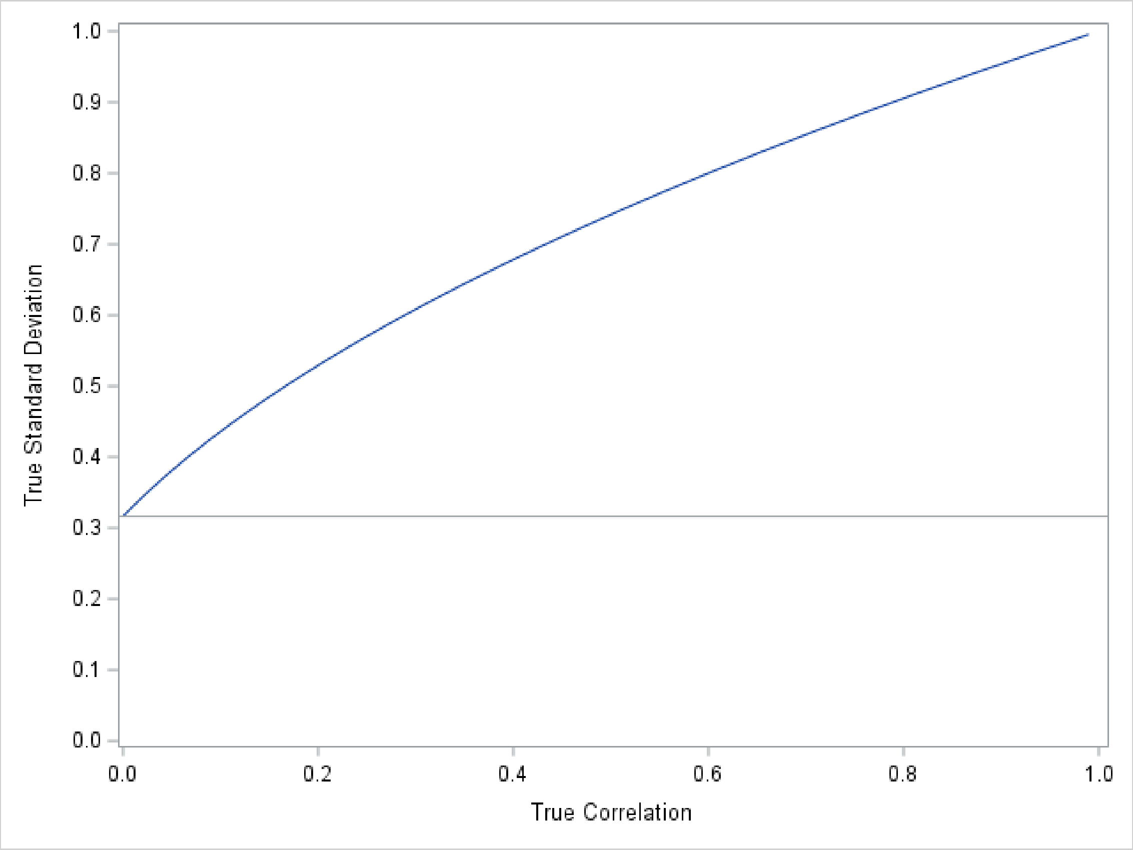
$$\begin{aligned} \Sigma &= \sigma^2 I_{n\times n} \\ Var(\hat{\beta}^{OLS}) &= [X^T\Sigma^{-1}X]^{-1} = \sigma^2(X^TX)^{-1} \end{aligned}$$

(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:

The SAS codes are as below



```
*****************************************************************
*
*  PROGRAM DESCRIPTION: Starter code for HW1, Question 2, Part D;
*
*---------------------------------------------------------------
*  JOB NAME:        HW1-Q2-PD.sas
*  LANGUAGE:        SAS, VERSION 9.4
*
*  NAME:            Matthew Psioda
*  DATE COMPLETE:   2020-01-18
*---------------------------------------------------------------
*
*  Modification History:
*
*  NAME:                          << Insert Name of Primary Programmer >>
*  DATE COMPLETE:                 << YYYY-MM-DD >>
*  DESCRIPTION OF MODIFICATION:   << Please insert 2-3 sentences >>
****************************************************************;

proc IML;
  rhoVec = t(do(0,0.99,0.01));                        ** set the correlation grid;
  plotDataIML = J(nrow(rhoVec),2,0);

  do r = 1 to nrow(rhoVec);
  rho         = rhoVec[r];                          ** extract the rth correlation;
  n           = 10;                                 ** set the sample
```

```
    size;
  x           = J(n,1,1);                          ** construct the
    design matrix;
  Sigma       = J(n,n,rho) + I(n)*(1-rho);         ** construct the
    covariance matrix;
  betaHatSD   = sqrt(inv(t(x)*inv(Sigma)*x));      ** compute the
    standard deviation;

  plotDataIML[r,] = rho||betaHatSD;                ** store current
    results in plotDataIML matrix;
  end;

  create plotData from plotDataIML[c={"rho" "stdErr"}]; ** write out a
    SAS dataset;
     append from plotDataIML;
  close plotData;
quit;


proc sgplot data = plotData;
  series x = rho y = stdErr;
  yaxis values=(0 to 1 by 0.1);
  refline %sysfunc(sqrt(1/10)) / axis=y;
  label rho    = 'True Correlation' stdErr = 'True Standard Deviation';
run;
```

(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

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

$$X = (1, ...1)^T, X^T X = 10$$

$$Var(\hat{\beta}^{OLS}) = (X^T X)^{-1} X^T \Sigma X [(X^T X)^{-1}]$$

$$= 10^{-1}(1, ...1) \begin{pmatrix} 1 & 1 & .. & 1 \\ 1 & 1 & .. & 1 \\ .. & .. & .. & \\ 1 & 1 & .. & 1 \end{pmatrix}_{10\times 10} (1, ...1)^T 10^{-1}$$

$$= 10^{-1}(10, ...10)(1, ...1)^T 10^{-1} = 1$$

When correlation coefficient is 1, we have the covariance is the same as variance, which shows that there is only one sample regression model. So the variance of the estimate $\hat{\beta}_{OLS}$ 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?

When we erroneously assume independent covariance matrix, $\Sigma = I$, then

$$\hat{\beta} = (X^TX)^{-1}X^TY = \bar{Y}$$
$$Var(\hat{\beta}) = (X^TX)^{-1}X^TIX[(X^TX)^{-1}] = (X^TX)^{-1} = \frac{1}{10}$$

So the critical value

$$z = \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}} = \bar{Y}\sqrt{10}$$

The rejection region with $1-\alpha\%$ CI is $z > Z_{1-\alpha/2}, z < Z_{-\alpha/2}$, under the independent covariance matrix.

While, when we assume correlation $\rho = \frac{1}{n-1}$, we have the corrected variance for $Var(\tilde{\beta})$.

$$\hat{\beta} = (X^TX)^{-1}X^TY = \bar{Y}$$
$$Var(\tilde{\beta}) = (X^TX)^{-1}X^T\Sigma X[(X^TX)^{-1}]$$
$$\Sigma = \begin{pmatrix} 1 & \rho & .. & \rho \\ \rho & 1 & .. & \rho \\ .. & .. & .. & \\ \rho & \rho.. & . & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{9} & .. & \frac{1}{9} \\ \frac{1}{9} & 1 & .. & \frac{1}{9} \\ .. & .. & .. & \\ \frac{1}{9} & \frac{1}{9}.. & . & 1 \end{pmatrix}$$
$$Var(\tilde{\beta}) = \frac{1}{10}(1, 1..1) \begin{pmatrix} 1 & \frac{1}{9} & .. & \frac{1}{9} \\ \frac{1}{9} & 1 & .. & \frac{1}{9} \\ .. & .. & .. & \\ \frac{1}{9} & \frac{1}{9}.. & . & 1 \end{pmatrix}$$
$$(1, 1..1)^T\frac{1}{10} = \frac{1}{5}$$

The actual variance of $\hat{\beta}$ is larger (1/5) than the variance assuming independent covariance (1/10). So the two sided rejection region will be smaller than the region based on incorrect covariance matrix. In other words, the type I error is inflated based on independent covariance matrix assumption.

The actual type I error rate based on independent covariance matrix assumption

$$p(x > \hat{\beta}|\beta = 0) = p(x < -\hat{\beta}|\beta = 0)$$
$$p = 2p(x > \hat{\beta}|\tilde{\beta}) = 2p(\bar{Y}\sqrt{10} > z_{1-\alpha/2})$$

We will use the below SAS codes to simulate the actual type I error rate, which is 0.16618.

The SAS codes are as below:

```
**********************************************************************
*
*  PROGRAM DESCRIPTION: Starter code for HW1, Question 2, Part E;
*
*-------------------------------------------------------------------
*  JOB NAME:       HW1-Q2-PE.sas
*  LANGUAGE:       SAS, VERSION 9.4
```

```
*
*  NAME:           Matthew Psioda
*  DATE COMPLETE:  2020-01-18
*-----------------------------------------------------------------
*
*  Modification History:
*
*  NAME:                          << Insert Name of Primary Programmer >>
*  DATE COMPLETE:                 << YYYY-MM-DD >>
*  DESCRIPTION OF MODIFICATION:   << Please insert 2-3 sentences >>
*******************************************************************;

proc IML;
 call randseed(1);

  n     = 10;
  rho   = 1 / (n-1);
  Sigma = J(n,n,rho) + I(n)*(1-rho);
  mu    = J(1,n,0);

  nSims = 50000;                         ** set number of simulations;
  y     = randNormal(nSims,mu,Sigma);    ** generate NSIMS random samples (row=sample);

  Z     = y[,:]/sqrt(1/n);               ** compute row means and standardize with incorrect SD;

  pVal  = 2*cdf('normal',-abs(z));       ** compute two-sided p-value based on known variance;

  rejRate = (pVal<0.05)[:];              ** compute proportion of samples rejected;

  print rejRate[l="Type I Error Rate"];

quit;
```

Generalize the type I error rate

$$\begin{aligned}\alpha_0 &= 2p(\bar{Y} > Z_{1-\alpha/2}SD(\hat{\beta})|\bar{Y} \sim N(0, Var(\tilde{\beta}))) \\ &= 1 - p(\frac{Z_{1-\alpha/2}sd(\hat{\beta})}{sd(\tilde{\beta})}) + p(-\frac{Z_{1-\alpha/2}sd(\hat{\beta})}{sd(\tilde{\beta})})\end{aligned}$$

The variance of $\tilde{\beta}$

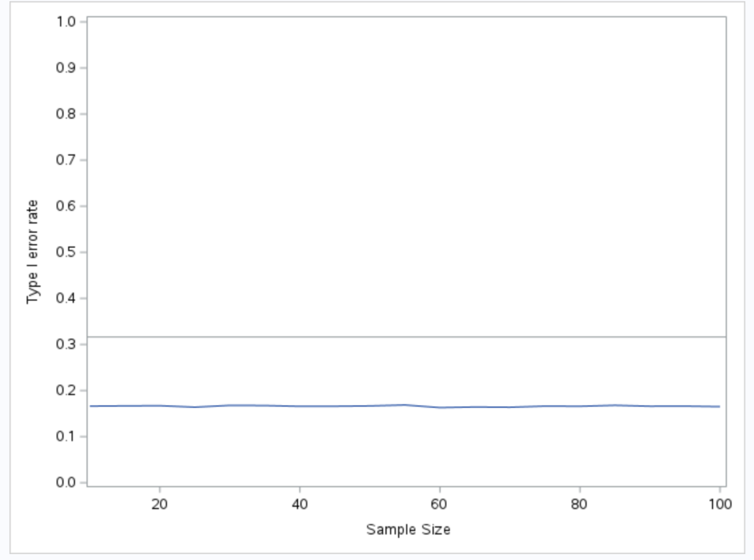
$$Var(\tilde{\beta}) = (X^TX)^{-1}X^T\Sigma X[(X^TX)^{-1}]$$

$$\Sigma = \begin{pmatrix} 1 & \rho & .. & \rho \\ \rho & 1 & .. & \rho \\ .. & .. & .. & \\ \rho & \rho.. & . & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{1}{n-1} & .. & \frac{1}{n-1} \\ \frac{1}{n-1} & 1 & .. & \frac{1}{n-1} \\ .. & .. & .. & \\ \frac{1}{n-1} & \frac{1}{n-1}.. & . & 1 \end{pmatrix}$$

$$Var(\tilde{\beta}) = \frac{1}{n}(1, 1..1)_{n\times 1} \begin{pmatrix} 1 & \frac{1}{n-1} & .. & \frac{1}{n-1} \\ \frac{1}{n-1} & 1 & .. & \frac{1}{n-1} \\ .. & .. & .. & \\ \frac{1}{n-1} & \frac{1}{n-1}.. & . & 1 \end{pmatrix} (1, 1..1)^T_{1\times n}\frac{1}{n} = \frac{2}{n}$$

So the generalized type I error rate is $1 - p(Z_{1-\alpha/2}\sqrt{2}) + p(-Z_{1-\alpha/2}\sqrt{2})$. We could see that the type I error rate doesn't change when sample size get larger. Also we use the simulated codes as above and changed the sample size n, the type I error rate is almost the same.

The plot of type I error rate vs. sample size as below:



The SAS codes for generating the above type I error rate figure as below:

```
proc IML;
 call randseed(1);
  nVec = t(do(10,100,5));                          ** set the sample size
    grid;
  plotDataIML = J(nrow(nVec),2,0);

  do r = 1 to nrow(nVec);
  n         = nVec[r];                      ** extract the rth sample size;
  rho    = 1 / (n-1);
  Sigma  = J(n,n,rho) + I(n)*(1-rho);
  mu     = J(1,n,0);
  nSims  = 50000;                            ** set number of simulations;
  y      = randNormal(nSims,mu,Sigma);      ** generate NSIMS random
    samples (row=sample);

  Z     = y[,:]/sqrt(1/n);              ** compute row means and standardize with incorrect SD;

  pVal  = 2*cdf('normal',-abs(z));      ** compute two-sided p-value based on known variance;

  rejRate = (pVal<0.05)[:];

  plotDataIML[r,] = n||rejRate;    ** store current results in plotDataIML matrix;
  end;

  create plotData from plotDataIML[c={"n" "rejRate"}]; ** write out a SAS dataset;
   append from plotDataIML;
  close plotData;
quit;


proc sgplot data = plotData;
  series x = n y = rejRate;
  yaxis values=(0 to 1 by 0.1);
  refline %sysfunc(sqrt(1/10)) / axis=y;
  label n    = 'Sample Size' rejRate = 'Type I error rate';
run;
```

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

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

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

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

(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.

From above, we have the variance of $\hat{\beta}_{hi,1}$, which is $(0,1)(X'_{hi}X_{hi})^{-1}X'_{hi}\Sigma_{J_{hi}\times J_{hi}}\Big[(X'_{hi}X_{hi})^{-1}X'_{hi}\Big]^T\begin{pmatrix}0\\1\end{pmatrix}$. The assumption of a common variance will require the X_{hi} and $\Sigma_{J_{hi}\times J_{hi}}$ are the same from different individuals. However, in longitudinal study, the design matrix X varies across individuals as they will have different measurement time. Also the covariance matrix between different measurements may be different. While in reality, the assumption does not hold due to the heterogeneity among individuals.

The sufficient condition for using t-test is to have i.i.d. observations among different groups, which requires the common variances of $\hat{\beta_{ji,1}}$. So we will need the design matrix and covariance matrix are the same for the two groups population.