HW 3

Bohao Zou
BST 224
University of California, Davis

May 16, 2020

Question 1

Question 1.(a)

Solution

$$\begin{aligned}\mathrm{Var}(Y_{ij}) &= \mathrm{Var}(U_{i1} + U_{i2} \times t_j + Z_{ij}) \\ &= \mathrm{Var}(U_{i1} + U_{i2} \times t_j) + \mathrm{Var}(Z_{ij}) \\ &= \mathrm{Var}(U_{i1}) + t_j^2 \times \mathrm{Var}(U_{i2}) + 2 \times t_j \times \mathrm{Cov}(U_{i1}, U_{i2}) + \mathrm{Var}(Z_{ij})\end{aligned}$$

Because $\mathrm{Var}(U_{i1}) = 12.25$, $\mathrm{Var}(U_{i2}) = 2.17$, $\mathrm{Cov}(U_{i1}, U_{i2}) = -1.52$ and $\mathrm{Var}(Z_{ij}) = 12.21$. We can know that :

$$\mathrm{Var}(Y_{ij}) = 2.17t_j^2 - 3.04t_j + 24.46 \tag{1}$$

The form of $\mathrm{Var}(Y_{ik})$ is as same as $\mathrm{Var}(Y_{ij})$. So, the $\mathrm{Var}(Y_{ik})$ is :

$$\mathrm{Var}(Y_{ik}) = 2.17t_k^2 - 3.04t_k + 24.46 \tag{2}$$

Question 1.(b)

Solution

$$\begin{aligned}\mathrm{Cov}(Y_{ij}, Y_{ik}) &= \mathrm{Cov}(U_{i1} + t_jU_{i2} + Z_{ij}, U_{i1} + t_kU_{i2} + Z_{ik}) && (3) \\ &= \mathrm{Var}(U_{i1}) + t_k \mathrm{Cov}(U_{i1}, U_{i2}) + t_j \mathrm{Cov}(U_{i1}, U_{i2}) + t_jt_k \mathrm{Var}(U_{i2}) && (4) \\ &= 2.17t_jt_k - 1.52(t_j + t_k) + 12.25 && (5)\end{aligned}$$

Question 1.(c)

Solution

$$\begin{aligned}\mathrm{Corr}(Y_{ij}, Y_{ik}) &= \frac{\mathrm{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\mathrm{Var}(Y_{ij})}\sqrt{\mathrm{Var}(Y_{ik})}} && (6) \\ &= \frac{2.17t_jt_k - 1.52(t_j + t_k) + 12.25}{\sqrt{2.17t_j^2 - 3.04t_j + 24.46}\sqrt{2.17t_k^2 - 3.04t_k + 24.46}} && (7)\end{aligned}$$

Question 1.(d)

Solution

$$\begin{aligned}\mathrm{Var}(Y_{ij}|U_i) &= \mathrm{Var}(Z_{ij}) && (8) \\ &= 12.21 && (9)\end{aligned}$$

The variance $\text{Var}(Y_{ij}|U_i)$ describes the variance of subject-specific model for the mean response on the i-th subject. This means that there is only one source of this variance. It comes from the factors which will effect i-th subject.

The variance $\text{Var}(Y_{ij})$ describes the variance of marginal model for the mean response on the over all subjects. This means that there are two source of this variance. Its come from the factors which will effect i-th subject and the different between individual.

Question 2

Question 2.(a)

Solution

For the question, i used two commands in R to complete this require.

$$ahead\_data\$realage = ahead\_data\$age + ahead\_data\$year$$
$$ahead\_data\$totword = ahead\_data\$immword + ahead\_data\$delword$$

Question 2.(b)

Solution

The model formula is

$$Y_{ij} = \beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + \beta_3 sex_i * age_{ij} + \beta_4 blks_{ij} + \cdots + U_i + Z_{ij} \tag{10}$$

The U_i and Z_{ij} are random variables and $U_i \sim \mathcal{N}(0, \nu^2)$, $Z_{ij} \sim \mathcal{N}(0, \tau^2)$. The U_i, Z_{ij} and X_i are all independent with each others.
The estimated values and the corresponding confidence intervals (95%) are showed in the table below.

Coefficient	Estimation	Confidence Interval 95%
β_1	-0.16	[-0.179 , -0.145]
β_2	0.80	[0.653 , 0.953]
β_3	-0.03	[-0.048 , -0.005]
β_4	-0.14	[-0.251 , -0.0213]

Table 1: *The table for displaying the estimated values and corresponding 95% cofidence interval.*

Interpretation

1. β_1 : It estimates the slope of *age* variable to the mean (overall) response variable (total words) for male subjects with the same difficulty situation.

2. β_2 : It estimates the difference of mean response variable (total words) at age 80 between male and female.

3. β_3 : It estimates the slope of *age* variable to the mean (overall) response variable (total words) for female subjects with the same difficulty situation.

4. β_4 : It estimates the slope of *blks* variable to the mean (overall) response variable (total words) for overall subjects.

Question 2.(c)

Solution

By using the formula $\beta_2 - 10 * \beta_3$, the estimated value for *sex effect for 70 year old* is 1.07. We know that the

$$\mathrm{Var}(\beta_2 - 10 * \beta_3) = \mathrm{Var}(\beta_2) + 100 * \mathrm{Var}(\beta_3) - 20 * \mathrm{Cov}(\beta_2, \beta_3)$$

So,the 95% confidence interval is $[0.821, 1.320]$.

Question 2.(d)

Solution

The null and alternative hypotheses are showed below.

$$H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_1 : one \quad of \quad \beta \quad is \quad not \quad 0$$

The test statistic is :

$$T_{stat} = (L\hat{\vec{\beta}})^T (L\hat{C}L^T)^{-1} (L\hat{\vec{\beta}}) \quad where \quad \hat{C} = \hat{\mathrm{Var}}(\hat{\vec{\beta}}|X)$$

L is a setting matrix which corresponding to the test coefficients. The $T_{stat} \sim F(DF_n, DF_d)$. The DF_n is the rank of L matrix. In general, it is hard to find the DF_d. So, we can use a χ^2 test to replace the F test because we have $\chi^2_{DF_n} = DF_n * F(DF_n, DF_d)$ in the situation that $DF_d \to \infty$. From above we can know that $T_{stat} * DF_n \sim \chi^2_{DF_n}$.

From this data set, we calculate the $T_{stat} = 100$. The $DF_n = 5$ because the rank of L is 5. The $p - value = 7.98 * 10^{-106}$. It can be consider as zero. We can get the conclusion that we need to reject the H_0 and accept H_1 and assert that the five physical function variables are jointly associated with total word recall.

Question 2.(e)

Solution

The estimated values and the corresponding confidence intervals (95%) are showed in the table below.

Coefficient	Estimation	Confidence Interval 95%
β_1	-0.16	[-0.179 , -0.143]
β_2	0.90	[0.754 , 1.044]
β_3	-0.04	[-0.066 , -0.020]
β_4	-0.26	[-0.391 , -0.122]

Table 2: *The table for displaying the estimated values and corresponding 95% cofidence interval for working correlation model.*

From the model formula of (10), we can interpret the coefficients of variable *age, sex, intersection of age and sex, blks.*

Interpretation

1. β_1 : It estimates the slope of *age* variable to the mean (overall) response variable (total words) for male subjects with the same difficulty situation.

2. β_2 : It estimates the difference of mean response variable (total words) at age 80 between male and female.

3. β_3 : It estimates the slope of *age* variable to the mean (overall) response variable (total words) for female subjects with the same difficulty situation.

4. β_4 : It estimates the slope of *blks* variable to the mean (overall) response variable (total words) for overall subjects.

We do not have to have the correlation structure exactly correct in order to obtain valid inferences. On the one hand it is because the mean model ($\vec{\beta}$) is our scientific interest but not the correlation model. Most of coefficient inferences are not rely on the correlation model. On the other hand, our model is based on have a correct model for V matrix. However, in general, we can not have a perfect correct V matrix. But GEE model with a simpler correlation structure and large samples can reduce the loss which we choose a wrong V matrix. Even though we selected a correct correlation structure. But it has a big probability that it is a complex correlation structure. This complex correlation matrix may not be as well estimated. In this data set, we have a large sample size and our scientific interest is on the mean model. The GEE model is suitable for our inference. So, we do not need to have a correlation structure exactly correct.

Question 3

Question 3.(a)

Solution

I used the following code to solve the (a) question.

$$
\begin{aligned}
&length(unique(birth\$id))\\
&length(which(birth\$birthorder == 1))\\
&length(which(birth\$birthorder == 2))\\
&length(which(birth\$birthorder == 3))\\
&length(which(birth\$birthorder == 4))\\
&length(which(birth\$birthorder == 5))\\
&birth\$momage = birth\$momage/10\\
&birth\$momage\_avg = birth\$momage\_avg/10\\
&birth\$momage\_dev = birth\$momage\_dev/10\\
&first\_birth = as.factor(as.numeric(birth\$birthorder == 1))\\
&birth\$first\_birth = first\_birth
\end{aligned}
$$

Question 3.(b)

Solution

The mixed effect model formula is :

$$brith\_wt_{ij} = \beta_0 + \beta_1 mom\_age_{ij} + \beta_2 first\_birth_{ij} + U_i + Z_{ij} \tag{11}$$

The U_i is the random effect and Z_{ij} respresents the residual of this model. $U_i \sim \mathcal{N}(0, \nu^2)$, $Z_{ij} \sim \mathcal{N}(0, \tau^2)$. The assumptions of U_i and Z_{ij} is that Z_{ij}'s are all independent of U_i and independent of X_i. U_i's are also independent of X_i.

The coefficients of fitted model (11) are showed in the table.

Coefficients	Value	Std.Error	t-value	p-value
β_0	2761.3	56.25	49.09	0.00
β_1	181.2	24.11	7.51	0.00
β_2	14.0	19.96	0.70	0.48
ν	354.1	Null	Null	Null
τ	434.3	Null	Null	Null

Table 3: *The table for displaying the fitted mix effect model of (11).*

Interpretation

1. β_1 : It estimates the slope of *mom age* variable to the mean response variable (*birth weight*) for the observations which are not first birth.

2. β_2 : It estimates the difference of mean response variable (*birth weight*) between the observations which are first birth and the observations which are not first birth but the variable *mom age* of those observations are same.

This model shows that there exists a positive association between the baby birth weight with maternal age. However, there dose not exist a significant association between the baby birth weight with first born effect.

Question 3.(c)

Solution

The fixed effect model formula is :

$$brith\_wt_{ij} = \beta_0 + \beta_1 mom\_age_{ij} + \beta_2 first\_birth_{ij} + U_i + Z_{ij} \tag{12}$$

The U_i is a fixed value and Z_{ij} represents the residual of this model. $Z_{ij} \sim \mathcal{N}(0, \tau^2)$. The difference between this fixed effect model and the mixed effect model is that in this fixed model U_i is not a random variable. It represents a fixed value which we do not know but need to estimate. In the mixed effect model, U_i represents a random variable. We need to estimate the variance of this random variable but not the value. In the fixed effect model, U_i can contain confounder and we do not make assumption that the U_i must independent with X_i.

The coefficients of fitted model (12) are showed in the table. I only present some of U_i in this table. Because the subjects in this data set is too large to display. The details are in my R code.

Coefficients	Value	Std.Error	t-value	p-value
β_0	3428.734	208.923	16.411	$2e^{-16}$
β_1	90.000	31.000	2.843	0.004
β_2	-28.912	22.213	-1.302	0.193
U_1	-254.841	274.773	-0.927	0.3537
U_2	-1128.400	274.417	-4.112	$4.01e^{-5}$
U_3	-668.810	275.035	-2.432	0.015

Table 4: *The table for displaying the fitted fixed effect model of (12).*

Interpretation

1. β_1 : It estimates the slope of variable (*mom age*) slope to the average (over U_i) response variable (*birth weight*) on the subject-specific (*i-th*) for the observations which are not first birth.

2. β_2 : It estimates difference of average (over U_i) response variable (*birth weight*) on the subject-specific (*i-th*) between the observations which are first birth with the observations which are not first birth but the variable *mom age* of those observations are same.

Why interpretations are differs in the solution of question 3 (b). This is because in the model (11), U_i is a random variable and it is a marginal model. The expectation of U_i is zero. So, we need to interpret those coefficients on the mean response variable.
However, in the model (12), U_i is a fixed value. It is different between different subjects. For controlling this effect, we need to interpret the coefficients on the conditional model, So, we need to interpret those coefficients on subject-specific average response variable.

This model shows that there exists a positive association between the baby birth weight and maternal age. However, here dose not exist a significant association between the baby birth weight and first born effect.

There is a explanations of the diverge of model (11) and model (12). It is that we give the different model with different assumptions.

1. In the fixed effect model, U_i is a fixed value, it can contain confounders which have correlative with X variables. So that, in the model $\vec{Y} = X\vec{\beta} + U_i + Z_{ij}, where\ Z_{ij} \sim \mathcal{N}(0, V)$, The matrix V will be $\sigma^2 I_{n_i}$, where I is the identical matrix. By using the formula $\hat{\vec{\beta}} = (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}\vec{y}$, we can know that $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{y}$ and the variance of $\vec{\beta}$ is $\mathrm{Var}(\hat{\vec{\beta}}|X) = (X^TX)^{-1}$.

2. In the mixed effect model, U_i is a random value, it must be independent with X variables. So that, in the model $\vec{Y} = X\vec{\beta} + U_i + Z_{ij}, where\ U_i + Z_{ij} \sim \mathcal{N}(0, V)$, The matrix V will not be a identical matrix, but with a different correlation structure in different assumptions. By using the formula $\hat{\vec{\beta}} = (X^T\hat{V}^{-1}X)^{-1}X^T\hat{V}^{-1}\vec{y}$, we can get the estimated $\hat{\vec{\beta}}$ coefficients variables. It is significantly different with the formula in fixed effect model. The variance of $\vec{\beta}$ is $\mathrm{Var}(\hat{\vec{\beta}}|X) = (X^T\hat{V}^{-1}X)^{-1}$.

Summary
Because of different model assumptions and different correlative structure, we can inference the relationship of U_i and X variables. Then conduct distinguish result of how to find estimated $\vec{\beta}$ and $\mathrm{Var}(\hat{\vec{\beta}}|X)$. Finally, using those statistic to get different coefficients, statistic inference and confidence interval.