Applied Statistical Modelling and Health Informatics

Institute of Psychiatry, Psychology and Neuroscience

King's College London

Assessment Coversheet

Submitting your work, either electronically or in hard copy, is regarded by your programme as confirmation that you have read, understood, and agreed to the Academic Misconduct Policy, the Non-Academic Misconduct Policy, and the Generative AI: student guidance on using AI tools with integrity and for effective learning.

You must complete all sections of this coversheet.

Candidate number	AF11817
Module title	Multilevel and Longitudinal Modelling
Module Code	7PAVMALM
Word count (Where a limit is specified).	3931
Please identify one aspect of your submission for which you would like specific feedback comments	

<u>Do not</u> include your name or K-number anywhere in your submission.

Student declaration

ANY FALSE DECLARATIONS WILL BE TREATED AS SERIOUS MISCONDUCT

By submitting this work (please tick to confirm):

1.	I confirm that I have not commissioned this work from a third party, and it represents **a genuine demonstration of my own skills and subject knowledge and is written using my own words**.	☒
2.	I understand that plagiarism, self-plagiarism, collusion, contract cheating and the inappropriate use of AI, **which includes copying directly from generative AI tools into submitted work**, are serious assessment offences, an allegation of which may lead to action being taken under the College's Misconduct regulations.	☒
3.	I understand that King's requires students to acknowledge the appropriate use of AI tools in assessments. If you are unsure how AI may be used during your studies, please check with your module and/or programme lead.	☒

Each statement below (1 through 4) represents an appropriate use of generative AI. Please tick **all** statements below that apply:

1.	Generative AI was not used at all for this assessment.	☒
2.	Generative AI was used to check spelling and grammar (but I confirm that the work submitted represents a genuine demonstration of my own skills and subject knowledge).	☒
3.	Generative AI was used to generate ideas or structure suggestions, for assistance with understanding core concepts, or other substantial foundational and preparatory activity. If ticked, briefly state which AI tool(s) were used, with links, and for which sections of the assignment AI was used. For example: "I used ChatGPT (https://chatgpt.com/) to generate ideas for the introduction, and a draft structure for the discussion."	☐
4.	Generate illustrative images (references should be provided for these within your main reference list). If ticked, state which AI tool(s) were used, with links, and briefly summarise how you used them.	☐

The use of generative AI to assist with writing code is only permitted if explicitly allowed by the module lead <u>in writing</u>.

5.	Generative AI was used to assist in writing code for the analysis reported in this assignment, including help with debugging. This box should only be ticked if approved by the module lead. If ticked, insert AI tool(s) and links and briefly summarise how you have used them.	☐

AF11817

January 2025

Section 1 Missing Pattern

In Table 1, a "1" represents the PANSS is observed and a "0" means missing. I got 46% of the data complete, 11% of the patients had no observed value at any point in time, and the remaining patterns showed missing PANSS scores at different points in time, with each pattern occurring less frequently (6% or less). Also, some data was monotonous missing, some data was random missing.

Table 1: Missing-value patterns for PANSS variables

Percent	**panss1**	**panss3**	**panss9**	**panss18**
46%	1	1	1	1
11%	0	0	0	0
6%	1	0	1	1
6%	1	1	0	1
6%	0	1	1	1
6%	1	1	0	0
5%	1	0	0	0
3%	1	1	1	0
3%	1	0	0	1
2%	0	0	0	1
2%	0	0	1	1
1%	0	1	0	0
<1%	0	1	0	1
<1%	0	1	1	0
<1%	1	0	1	0

Section 2 Summarise PANSS remission

It could be seen from the table 2 that the remission rates of PANSS in the intervention group and the control group at different time points are as follows:

Table 2: Summary of PANSS Remission Scores by Intervention Group

Interven	base	6weeks	3months	9months	18months
Control	0.17	0.59	0.63	0.68	0.61
Intervention	0.14	0.63	0.69	0.75	0.74
Total	0.15	0.62	0.67	0.73	0.70

At baseline, the response rate in the intervention group (14%) was slightly lower than in the usual care group (17%). Over time, remission rates were higher in the intervention group than in the usual care group. At 18 months, the response rate was 74% in the intervention group, compared with 61 percent in the usual care group. This suggests that intervention may be beneficial for remission.

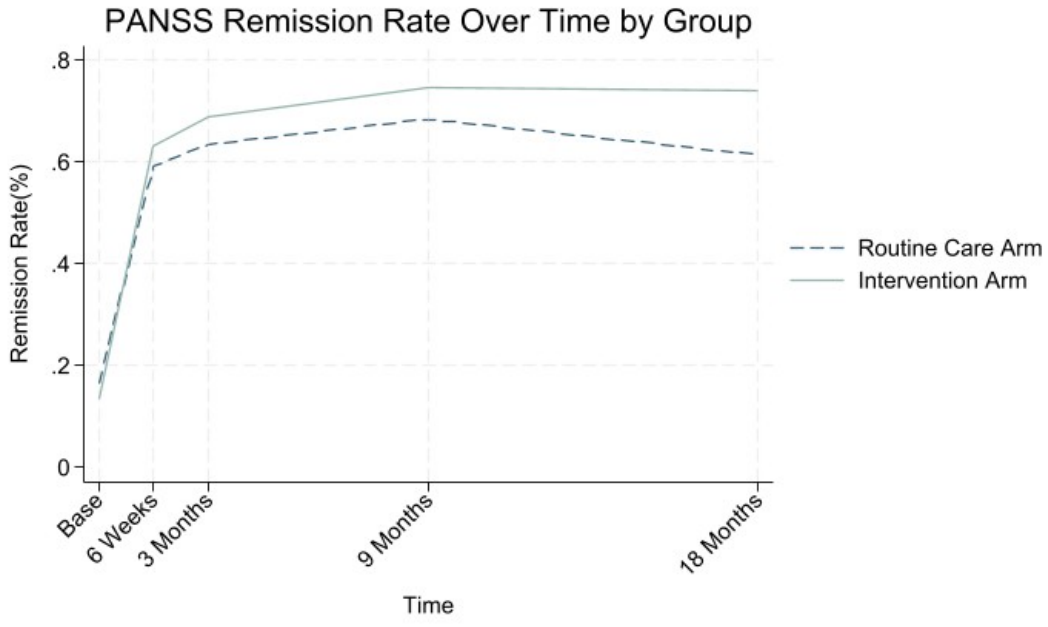

Figure 1: PANSS Remission Rate Over Time By Group

It could also be seen from Figure 1 that after 6 weeks, the remission rate of the intervention group exceeded that of the conventional care group, and the remission rate of both

groups gradually increased, but the remission effect of the conventional care group was better than that of the control group. The remission rate also decreased in both groups after 9 months, but the usual care group showed a significant decrease in fluctuation after 9 months. In general, remission was better in the intervention group than in the control group, especiall after 6 weeks. From a statistical point of view, although the relevant intervention has shown some effectiveness, the reliability and representativeness of the results are still limited and further modeling and analysis are still needed.

Section 3 Longitudinal Mean PANSS Profiles

To illustrate the longitudinal changes in PANSS mean scores between the two groups, I visualized them using line graphs and box graphs. As shown in Figure 2, the mean baseline values for the intervention group and the control group were 88.59 and 86.96, respectively. PANSS scores declined over time in both groups. However, the reduction was significantly greater in the combined intervention group compared to the usual care group. At both the 3 and 9 month time points, the scores in the intervention group were significantly lower than those in the usual care group. According to Figure 3, the mean score of the intervention group (60.18) was significantly lower than that of the control group (66.40), especially in the 18th month. This suggests that combined intervention may be more effective than conventional care.

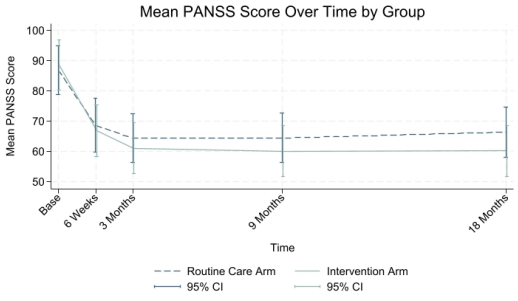

Figure 2: Mean Panss over Time

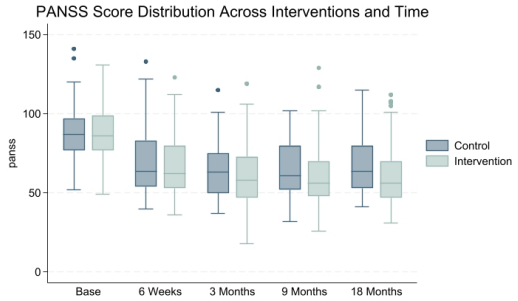

Figure 3: PANSS Scrore distribution over time

From figure 4 in the Intervention group, the decline appeared to be more pronounced and more stable.

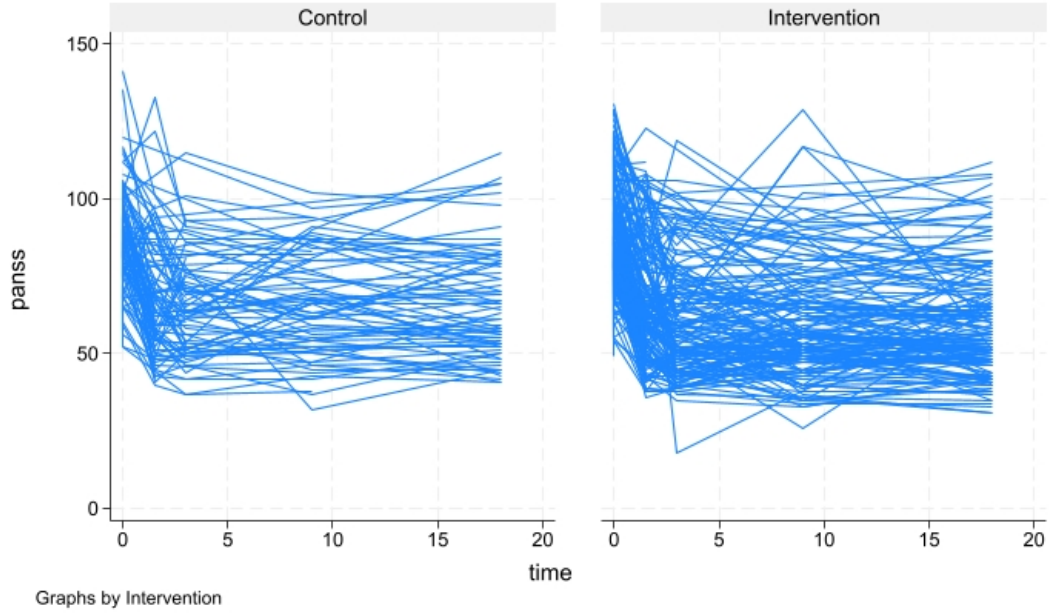

Figure 4: Individual PANSS Score over Time By Group

Section 4 GLMM Estimation of Treatment Effect on PANSS Scores at 18 Months

Because each patient was measured multiple times, the observations were not independent of each other, and their PANSS scores were correlated over time. Although the individual trajectories were roughly linear, the slope and intercept varied between patients (Figure 4) and therefore we cannot simply use:

$$y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{1}$$

Considering that the intercept is likely to be disturbed by some random factors, I first established a random intercept model with time variables:

$$y_{ij} = (\beta_0 + u_i) + \beta_1 \cdot \text{Time}_{ij} + \varepsilon_{ij}, \quad u_i \sim \mathcal{N}(0, \sigma_u^2), \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

In particular, the time variable was centralised with an 18-month baseline.My aim was to find a proper time scalling. This helps to explain the coefficient and avoid the influence on the interpretation when the starting point of the time variable is 0 months. Logarithmic transformation of dependent variable PANSS is helpful to improve model fitting because PANSS data is biased to the left.

Table 3: Model 1_{test1}: Random Intercept Mixed-Effects Model with Centralized Time

Fixed Effects	**Coef.**	**Std. Err.**	**z**	**95% CI**
month_change	-0.00154	0.00043	-3.59	[-0.00238, -0.00070]
Intercept	1.77267	0.00779	227.63	[1.75741, 1.78793]
Random Effects (Variance Components)				
Intercept (idnumber)	0.00870			[0.00702, 0.01079]
Residual	0.00676			[0.00604, 0.00757]
Model Summary				
Number of observations				877
Number of groups (idnumber)				274
Log likelihood				728.54
Wald $\chi^2(1)$				12.89 (p = 0.0003)

Models with a centralized time variable were statistically significant (Wald χ^2 = 12.89, $P < 0.001$), i.e., the coefficient of time was significantly different from zero. As the model included only one centralized independent variable (time), its coefficient was tested using a t-test ($t = -3.59$, $P < 0.001$).

And then I tried to include some more variables. According to the question, I consider ”treatment group” as the ”confounding factor” and other variables as background covariables to fit the random intercept model. What’s more, I choose to think of ”centre” and ”therapist” as higher-level clustering variables, representing level-3 clusters in the data, rather than fixed covariates.

According to the Table 4, eight variables were newly included in the Model 1_{test2}, including the baseline PANSS score, which accounts for the different initial conditions of patients before treatment. Model 1_{test2} was statistically significant (Wald $\chi^2 = 162.57$, $p < 0.001$). Compared with the previous model, the variance of the random intercept has decreased, suggesting that the newly added variables contribute to explaining the differences between subjects in the intercepts.

However, the individual coefficients for gender, drug abuse, years of education, age at study entry, number of hospital admissions, were not statistically significant, indicating

that these variables are not associated with changing in the dependent variable.

Table 4: Non-significant predictors in the Model 1_{test2}

Variable	**Coefficient**	**Std. Err.**	**z**	**95% CI**
Male (1 = male)	0.00941	0.01184	0.79	[-0.01380, 0.03261]
Substance misuse: weekly	0.00938	0.02017	0.47	[-0.03015, 0.04892]
Substance misuse: daily	-0.00760	0.01839	-0.41	[-0.04365, 0.02844]
Substance misuse: unknown	-0.00584	0.01777	-0.33	[-0.04066, 0.02898]
Substance misuse: none	0.01327	0.01518	0.87	[-0.01649, 0.04303]
Years of education	-0.00056	0.00232	-0.24	[-0.00510, 0.00397]
Episode (2nd episode)	0.01305	0.01462	0.89	[-0.01561, 0.04170]
Age at entry	-0.00005	0.00054	-0.09	[-0.00110, 0.00100]

From the perspective of simplifying model, the above 5 variables are removed before building Model 1_{test3}.

Table 5: Result of Model 1_{test3}

Fixed Effects	**Coef.**	**Std. Err.**	**z**	**95% CI**
month_change	-0.00144	0.00043	-3.35	[-0.00228, -0.00060]
logbase	0.64398	0.06408	10.05	[0.51839, 0.76957]
Intervention	-0.03441	0.01090	-3.16	[-0.05578, -0.01303]
logdup	0.04993	0.00892	5.60	[0.03244, 0.06741]
Intercept	0.48928	0.12449	3.93	[0.24528, 0.73327]
Random Effects (Variance Components)				
Var(Intercept, idnumber)	0.00466			[0.00359, 0.00603]
Var(Residual)	0.00682			[0.00609, 0.00764]
Model Summary				
Number of observations				867
Number of groups (idnumber)				271
Log likelihood				779.087
Wald $\chi^2(4)$				163.86 ($p < 0.001$)

At Model 1_{test3}, it was statistically significant (Wald $\chi^2 = 163.86$, $p < 0.001$). However, the residual variance was not better explained, which meant it increased lightly. This illustrates that Model 1_{test3} was insufficient to explain the.variation among the intra-individual.

Therefore, it is reasonable to consider a more complex random-effects structure by allowing for varying slopes. The random intercept-slope model fitted next assumes that individual patients not only differ in their intercepts (i.e., baseline levels), but also in the slopes (i.e., rates of change over time)

Before modelling, I checked the correlation matrix(Table 6) of logpanss and found that the correlation gradually decreased over time. However, the correlation between logpanss_1 and logpanss_9 (0.44) was lower than that between logpanss_1 and logpanss_{18} (0.46), which indicated that the model needs a more flexible covariance structure considering that the covariance structure of the composite symmetry was not fully fit. In my view, the correlation changes from point to point are complex and I try to use an "Unstructured" structure that allows for different correlations from point to point without making any assumptions.

Table 6: Correlation Matrix of `logpanss` at Different Time Points

	logpanss_1	logpanss_3	logpanss_9	logpanss_18
logpanss_1	1.0000			
logpanss_3	0.6639	1.0000		
logpanss_9	0.4452	0.6000	1.0000	
logpanss_18	0.4558	0.5533	0.6447	1.0000

The fitted random intercept-slope model (Model 1) was statistically significant (Wald χ^2 = 162.84, $p < 0.001$), and all variables were significant. Especially, due to the lack of statistical significance, the interaction term (Intervention $\times$ Time) was not considered further in the analysis.

Comparing Model 1 and Model 1_{test3}, the fixed effects are the same but the random effects are different. Table 7 showed that there was a significant difference between the two models ($p = 0.02 < 0.05$), indicating that model 1 was significantly superior to Model 1_{test3}. At the same time, from the comparison of AIC and BIC, the AIC

of model1 was -1547.61, which was lower than that of the model with only Model 2 (-1544.18), indicating that Model 1 had better fitting effect. However, although the relatively high BIC value (-1504.72) indicates that model 1 is more complex, the fit improvement brought about by this increased complexity led us to choose model 1 as the final model.

Table 7: Model 1 vs.Model 1_{test3}

Model	**N**	**Log Likelihood**	**df**	**AIC**	**BIC**
Model 1_{test3}	867	779.09	7	-1544.18	-1510.82
model1	867	782.80	9	-1547.61	-1504.72

Likelihood-ratio test: LR $\chi^2(2) = 7.44$, $p = 0.0243$

Finally, we got the Model 1:

$$\begin{aligned}\text{logPANSS}_{ij} = {} & \beta_0 + \beta_1 \cdot (\text{Time}_{ij} - 18) + \beta_2 \cdot \text{Intervention}_i + \beta_3 \cdot \text{logBase}_i + \beta_4 \cdot \text{logDup}_i \\ & + u_{0i} + u_{1i} \cdot (\text{Time}_{ij} - 18) + \varepsilon_{ij} \end{aligned} \tag{3}$$

According to the table 8, the effect of treatment was reflected in the coefficient of the variable `interven`, which represented the average difference in the log-transformed PANSS score between the intervention and control groups, controlling for time and other covariates. Among them, the PANSS score in the intervention group was about 3.30% lower than that in the control group (95% CI: -5.35%, -1.22%, $p < 0.001$), indicating a significant effect of the intervention.

Meanwhile, the predicted marginal mean showed that PANSS scores in the intervention and control groups were 1.76 and 1.80 (Table 9), respectively, on a logarithmic scale at 18 months, further indicating that PANSS levels in the intervention group were lower than those in the control group. Since the difference between the groups was statistically significant ($p < 0.001$), this suggests that the intervention still maintained a significant therapeutic advantage at 18 months.

Table 8: Mixed-Effects Model with Random Intercepts and Random Slopes (Unstructured Covariance)

Fixed Effects	**Coef.**	**Std. Err.**	**z**	**95% CI**
month_change	-0.00144	0.00047	-3.05	[-0.00236, -0.00051]
Intervention	-0.03365	0.01089	-3.09	[-0.05499, -0.01231]
logbase	0.64504	0.06397	10.08	[0.51966, 0.77042]
logdup	0.05017	0.00891	5.63	[0.03271, 0.06762]
Intercept	0.48634	0.12431	3.91	[0.24269, 0.72999]
Random Effects (Unstructured Covariance Matrix)				
Var(Intercept, idnumber)	0.00646			[0.00453, 0.00921]
Var(month_change)	0.0000137			[0.00000628, 0.00003]
Cov(Intercept, month_change)	0.00015			[0.00002, 0.00028]
Residual variance	0.00601			[0.00522, 0.00692]
Model Summary				
Number of observations				867
Number of groups (idnumber)				271
Log likelihood				782.805
Wald $\chi^2(4)$				162.84 ($p < 0.001$)

Table 9: Estimated Marginal Means of logPANSS by Group

Group	**Margin**	**Std. Err.**	**z**	**95% CI**
Control	1.80	0.01	174.51	[1.78, 1.82]
Intervention	1.76	0.01	221.75	[1.75, 1.78]

When using maximum likelihood estimation, it is commonly assumed that the residuals follow a normal distribution to simplify the likelihood function and facilitate parameter estimation. To better satisfy this assumption, a logarithmic transformation was applied to the dependent variable, aiming to make the residuals more normally distributed. However, this normality assumption may not always reasonable.

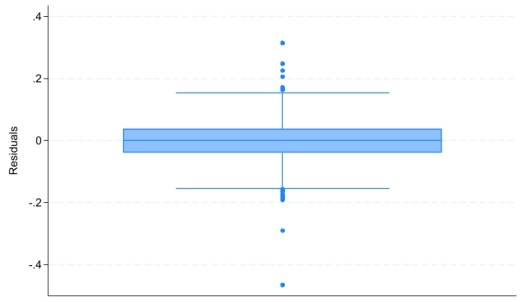

Figure 5: Mean Panss over Time

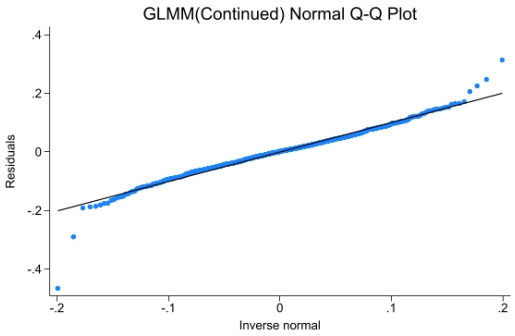

Figure 6: PANSS Scrore distribution over time

The residual distribution is roughly normal (Figure 5), and the points are well distributed near the diagonal, indicating that the residual is in line with the normal assumption and does not show obvious skewness or outliers. According to Figure 6, the residual is basically symmetrical and there is no heteroscedasticity.

Section 5 GLMM Estimation of Treatment Effect on PANSS Remission at 9 Months

When the result is a binary variable, the background covariate refers to the variable selected by the final model of session 4, and the time variable is centralized based on the baseline of 9 months.First, I firted the random intercept model(Model 2_{test1}), generalized logistic model of random intercept slope(Model 2).

Compared with Model 2_{test1}, Model 2 significantly improved the model fit (LR $\chi^2 = 23.63$, $p < 0.001$). In addition, the AIC (888.47) and BIC (921.83) values of Model 2 were lower than those of Model 2_{test1} (AIC = 896.17, BIC = 924.76), further supporting model 2 was better. The interaction effect (`interven × time_center`) was not statistically significant, so it won't consider in the final model.

Table 10: Model 2 vs. Model 2_{test1}

Model	N	Log Likelihood	df	AIC	BIC
Model 2_{test1}	867	-442.08	6	896.17	924.76
model2	867	-437.24	7	888.47	921.83

Likelihood-ratio test: $LR\ \chi^2(1) = 9.70,\ p = 0.0018$

The final model is:

$$\begin{aligned}\text{logit}\left(P(y_{ij}=1)\right) &= \beta_0 + \beta_1 \cdot (Time-9) + \beta_2 \cdot \text{Intervention}_i \\ &\quad + \beta_3 \cdot \text{BaseRemin}_i + \beta_4 \cdot \log(\text{Dup}_i) + u_{0i} + u_{1i} \cdot (Time-9) \quad (4)\end{aligned}$$

Since I centralised the monthly variables, the therapeutic effect in the fixed effect can be directly explained by the difference between the intervention and control groups at 9 months(Table 11). At the baseline time point (9 months), the probability of PANSS remission in the intervention group was about 2.61 times that of the control group (OR = 2.61, 95% CI: 1.04,6.53; p = 0.04), which was statistically significant. In addition, with each additional month over time, the odds of achieving remission increased significantly (OR = 1.05, 95% CI: 1.01,1.10; p = 0.02), that is, the odds of remission increased by about 5.4% per month, suggesting a trend of improvement over time.

Table 11: Result of Model 2

Variable	**Odds Ratio**	**Std. Err.**	**z**	**95% CI**
month_cen	1.054	0.0244	2.25	[1.007, 1.103]
Intervention	2.606	1.2241	2.04	[1.038, 6.543]
baseremin	22.365	19.565	3.55	[4.026, 124.225]
logdup	0.129	0.0562	-4.70	[0.055, 0.303]
Intercept	19.604	12.635	4.62	[5.543, 69.337]
Random Effects (idnumber)				
Var(month_cen)	0.0284 [0.0105, 0.0772]			
Var(Intercept)	8.0427 [4.4887, 14.4106]			
Model Summary				
Number of observations				867
Number of groups (idnumber)				271
Log likelihood				-437.236
Wald $\chi^2(4)$				31.50 ($p < 0.001$)

For this model, the residual of its random effects should satisfy the multivariate normal distribution. The empirical Bayesian estimates of random intercept and random

slope show an approximately elliptical distribution with a symmetric center around zero(Figure 7), and the correlation between the two is very low (-0.003), indicating that the random effect satisfies the modeling assumption of the multivariate normal distribution.

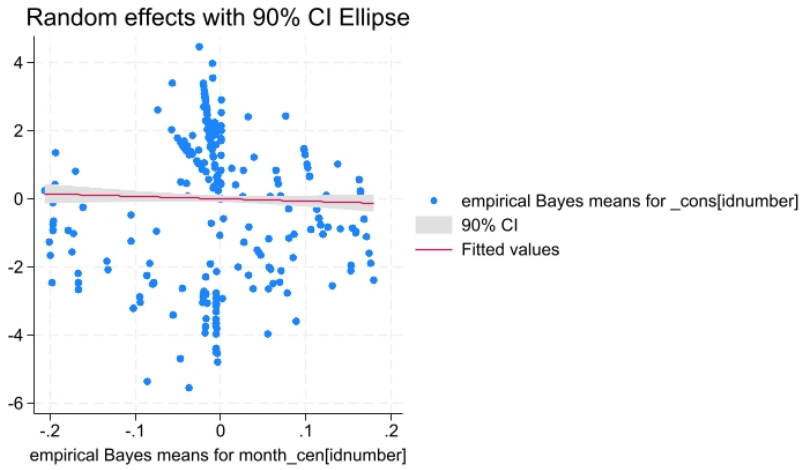

Figure 7: Joint Distribution of Empirical Bayes Estimates for Random Effects

Section 6 Treatment Effect Estimation on PANSS Using GEE

In session 6, the selection of covariables is the same with the above, and the time variable is unified as "week" as the unit. I adjusted the time variable by unifying units. Although equidistant time could be constructed by recoding to use an autoregressive structure $(AR(1))$ in order to varying intervals of time points, this structure was not considered because it would diminish the practical clinical significance of time. But in this data, considering the structure of $AR(1)$ is meaningful, as the analysis before, with the increase of time, the score will decrease and the mitigation score will increase, indicating that there may be autocorrelation between the data.

The exchangeable and unstructured work-related structures were compared in the GEE model, and I chose QIC value to estimate them (QIC serves the same purpose as AIC). Table 12 showed that the exchangeable structure fitted the best, and finally the structure was selected for analysis.

Table 12: Comparison of Correlation Structures Based on QIC and QIC_u

Correlation Structure	**QIC**	**QIC_u**
Unstructured	218354.11	218347.84
Exchangeable	217831.87	217824.97

According to exchangeable structure, I fitted Model 3. As the interaction term was not statistically significant, it would be excluded from Model 3.

Table 13: GEE Model Results with Exchangeable Correlation Structure

Variable	**Coefficient**	**Std. Err.**	**z**	**p-value**	**95% CI**
logDUP	7.70	1.37	5.62	<0.001	[5.12, 10.28]
Baseline PANSS	0.48	0.04	10.79	<0.001	[0.39, 0.57]
Intervention	-4.72	1.68	-2.82	0.003	[-7.87, -1.56]
Time (centered, months)	-0.21	0.07	-3.01	<0.001	[-0.34, -0.08]
Intercept	14.12	4.16	3.39	<0.001	[5.44, 22.80]

From Table 13, both duration of untreated psychosis (logDUP) and baseline PANSS score were significantly positively associated with PANSS during follow-up ($p < 0.001$). PANSS scores in the intervention group were significantly lower than those in the control group ($\beta = -4.72$, $p = 0.003$), indicating that the intervention had a significant effect on relieving psychiatric symptoms. In addition, time (centered at 18 months) was significantly negatively associated with PANSS ($p < 0.001$), indicating an overall improvement in symptoms over time. Taken together, the intervention has a clear therapeutic advantage in the overall population.

In order to test the robustness of the model estimation results, I adopted robust standard error. Through Table14, the conclusions of the model do not change, what's more, the regression coefficient of the core explanatory variables (such as the intervention coefficient) changed from -4.72 to -4.71, but the magnitude of the change was almost negligible and the level of statistical significance remained stable. In addition, the confidence interval range of the control variables did not significantly widen or narrow, indicating that the model was stable.

Table 14: Model 3 Results with VCE

Variable	Coefficient	Std. Err.	z	p-value	95% CI
logDUP	7.70	1.37	5.61	<0.001	[5.01, 10.38]
Baseline PANSS	0.48	0.04	10.79	<0.001	[0.39, 0.57]
Intervention	-4.71	1.68	-2.81	0.005	[-7.99, -1.43]
Time (weeks)	-0.054	0.017	-3.12	0.002	[-0.088, -0.020]
Intercept	17.94	4.09	4.39	<0.001	[9.93, 25.95]

Model 3 used an exchangeable structure that assumed the same correlation between observations for the same patient at different points in time. As can be seen from the table 15, the estimated correlation coefficient was about 0.40, indicating a moderate degree of agreement between PANSS measurements within individuals. This result supported the rationality of setting exchangeable structure.

Table 15: Estimated Within-Subject Correlation Matrix (Exchangeable Structure)

	c1	c2	c3	c4
r1	1			
r2	0.40	1		
r3	0.40	0.40	1	
r4	0.40	0.40	0.40	1

In terms of sensitivity analysis, I conducted sensitivity analysis from three dimensions: time variable setting, dependent variable conversion and whether interaction terms were included or not. First, in terms of Time variable processing, time was incorporated into the model in original form and centralized form (time-18) respectively. The results showed that the direction of intervention effect was the same under the two Settings, and the time trend was also significant, indicating that the model had a good robustness for time expression. In addition, the model after logarithmic transformation was constructed showed that the intervention effect was statistically significant (p = 0.003), and the intervention effect was relatively stable. Finally, in order to test the sensitivity of the model to the setting of interaction terms, without the interaction terms of ”intervention

× time" was built. The results showed that after the addition of the interaction term (P=0.193), the main effect of the intervention was weak (p = 0.108), while the direction and significance of other covariables (such as baseline variables) remained consistent. In summary, although the model has some sensitivity to the setting of interaction terms, the intervention effect is more robust and significant in the model with more concise structure.

AS for assumption, I assumed that missing data were either missing at random (MAR) which is associated with covariates, or missing completely at random (MCAR), which is a common assumption under the GEE framework. This assumption would be valid on the premise that there are omissions, and stata would remove the omissions during the model building process. In the selection of variance function, Gaussian distribution family and identity link function are used to apply to continuous PANSS score variables.

Section 7 Treatment Effect Estimation on PANSS Remission Using GEE

Similar to session 6, I choose similar covariates. However, panss release and release at baseline time became dichotomy in terms of dependent and baseline variables. Under the premise of the unified standard of time week, select the correlation structure.

From Table 16, I got Exchangeable structure was better, because its QIC and QIC_u were lower than that in Unstructured structure. IN the model 4, I adopted Exchangeable structure.

Table 16: Comparison of GEE Correlation Structures Based on QIC and QIC_u

Correlation Structure	**QIC**	**QIC_u**
Exchangeable	1014.12	1007.43
Unstructured	1014.36	1008.08

The results of Model 4 are as follows in Table 17:

Table 17: Model 4 Results for PANSS Remission (Logit Link, Binomial Family)

Variable	OR	Std. Err.	z	p-value	95% CI
Time (weeks)	1.01	0.00	2.44	0.02	[1.00, 1.01]
Intervention	1.57	0.35	2.05	0.04	[1.02, 2.42]
Baseline PANSS	0.95	0.01	-7.87	0.00	[0.93, 0.96]
DUP	1.00	0.00	-2.98	0.00	[0.99, 1.00]
Intercept	203.40	133.41	8.10	0.00	[56.24, 735.65]

After baseline correction, and duration of disease treatment, the remission rate of PANSS in the intervention group was 1.57 times that of the control group (OR = 1.57, p = 0.04), suggesting that the intervention had a significant therapeutic effect. Also, I found that time effect showed positive correlated(OR = 1.01), we could infer that patient would get better with the development of time.

Table 18: Model 4 with Robust Standard Errors (Binomial, Logit Link)

Variable	OR	Std. Err.	z	p-value	95% CI
Time (weeks)	1.01	0.00	2.19	0.03	[1.00, 1.01]
Intervention	1.57	0.35	2.02	0.04	[1.01, 2.44]
Baseline PANSS	0.95	0.01	-7.77	0.00	[0.93, 0.96]
DUP	1.00	0.00	-2.12	0.03	[0.99, 1.00]
Intercept	203.40	138.62	7.80	0.00	[53.49, 773.50]

For robust analysis, I also introduced vce, and obtained from the Table that the odds ratio (OR) of each variable remained unchanged, but the standard error increased slightly, and the p-value became more conservative. However, the effects were still statistically significant. This suggests that the model estimates are robust even when the job-related structure may be set inaccurately.

Table 19: Estimated Working Correlation Matrix (Exchangeable Structure)

	c1	**c2**	**c3**	**c4**
r1	1.00			
r2	0.31	1.000		
r3	0.31	0.31	1.000	
r4	0.31	0.31	0.31	1.000

The working correlation matrix was the estimation result of Exchangeable structure set in Model 4 with Robust standard error, and the correlation coefficient was 0.306. This showed that there were some autocorrelations within individuals.

Next, sensitivity analysis is carried out to test the centralization of the Time variable (Time-9) and the addition of interaction terms. After time centralization, the model was significant overall (Wald χ^2 = 72.30, P¡0.001), and all variables were statistically significant, showing that the treatment effect of the intervention group was better than that of the control group. This shows that model 4 is relatively stable. Then, adding the interaction term(Interven $\times$ Time), although the overall model is significant, the time variable ($P = 0.71$; 95% CI: 1,1.01), intervention variables($P = 0.25$; 95% CI 0.80,2.39), interaction terms($P = 0.31$; 95% CI: 1,1.02) were not significant. I think it is because there is no significant statistical difference between intervention effects at different time points. In addition, the introduction of interaction terms may weaken the main effect expression of the model, which also indicates that simplifying the model structure can more stably reveal the overall effect of the intervention.

I have used the binomial family and the logit link function. This was suitable for binary outcome.

Section 8 Comparison and Summarisation

Model Findings and Differences

The interaction terms of the time variable and the intervention variable were not statistically significant in either the GLMM or GEE model, continuous or discrete. This

indicates that the intervention effect is relatively stable, even if the intervention effect is significant, the rapid decline of PANSS score after treatment cannot be achieved in the analysis of this chapter.

At the same time, the intervention group showed an improvement trend. In the mixed-effect model, the intervention effect was statistically significant at the individual level ($p < 0.05$), while in the GEE model (Model 3 with Robust Standard error), the continuous type score was significant but the subtype result was only marginal ($p = 0.05$).

When continuous PANSS was the dependent variable, the regression coefficient of the intervention group in GLMM was negative (coefficient of -0.003365 in model 1), and the coefficient of GEE was -4.72 (Model 2). The intervention could significantly reduce PANSS score, reflecting the relief of psychiatric symptoms. In the binary PANSS mitigation model, the intervention OR was greater than 1, for instance, the GLMM OR value was 2.606 (Model 3); It was 1.57 in GEE (model 4), indicating that the intervention significantly improved the likelihood of achieving a remission state. The two types of models are in the same direction and both support the effectiveness of the intervention.

The assumption of missing data

GLMM requires MAR and GEE requires a stronger MCAR or covariation-dependent absence. In the main analysis mentioned above in this study (including GEE analysis), the observed value at each time point was 867.

Marginal vs Conditional Effects

The fixed effect of GEE is directly marginal effect. The fixed effect of the random effects model is the conditional effect. Because marginal effects do not take into account individual level differences, they are usually smaller than conditioned effect estimates. For example, the OR of the time coefficient in Model 2 (GLMM) was 1.05, while it was 1.02 in Model 4 when the time coefficient was cantered.

Robustness and limitations of results

For the robustness of the model, I chose time variable centralization, dependent variable logization, interaction term, and work correlation matrix analysis. The results were basically similar, and the significance of the main variables was basically stable, indicating that the model conclusions had certain robustness.

However, there was an imbalance in sample size between the control group and the intervention group, and the missing data was not processed in this study, which may introduce estimation bias. In addition, the interaction terms between intervention and time were not statistically significant under various models, which limited the further exploration of the trend of intervention effects over time.

Section 9 Exploring Additional Sources of Clustering and Model Extensions

I think patients come from different centers and different therapists, making up a potential cluster structure. However, due to the small number of centers, they are not suitable for modeling as random effects, and are more suitable for fixed effects. At the same time, the therapist variable is not suitable for the model, because there are more data missing, and the intervention group and the therapist are highly confused, unable to distinguish the independent contribution of the two, and the intervention effect may be affected by the therapist's level.

```
mixed logpanss c.month i.interven logbase c.logdup ///
      || centre: , ///
      || idnumber: month_change, cov(unstructured)
```

Listing 1: Stata mixed model syntax

Table 20: Description of Parameters in the Mixed-Effects Model

Parameter	**Type**	**Description**
`c.month`	Fixed effect	Effect of time (change in log(PANSS) per time unit).
`i.interven`	Fixed effect	Effect of intervention group compared to control group.
`logbase`	Fixed effect	Baseline log(PANSS) score to adjust for initial severity.
`c.logdup`	Fixed effect	Effect of covariate `logdup` (e.g., dosage or comorbidity).
`centre`	Random effect	Random intercept accounting for center-level variation.
`idnumber`	Random effect	Subject-specific random intercept and time slope.
`cov(unstr)`	Covariance structure	Unstructured covariance matrix for intercept and slope, allowing correlation.

If these models are used, the number of centers is small, which leads to unstable estimates of random effects, resulting in large estimates of the model variance. Moreover, adding multiple random effects increases the degree of freedom of the model and may lead to fitting difficulties (such as non-convergence). As for the treatment effect concerned in this case, the above issues are more concerned with the treatment effect and time effect. I prefer to take centre as a third-order variable, so I think its inclusion in the model will not help to further explain the results. In addition, the current number of individuals assessed per patient excluding baseline is 4, and the variation between patients is not easily stabilized.

Section 10 Dropout Analysis in Monotonic Missing-ness Subsample

Primarily, I screened out the patients who met the monotony loss, and obtained 222 patients who met the monotony loss. Data that has baseline data and falls off from week 6, I think, is also monotonous

Table 21: Missing-value Patterns of PANSS Measurements

Pattern	panss1	panss3	panss9	panss18	Frequency
1	1	1	1	1	143
2	1	1	1	0	10
3	1	1	0	0	18
4	1	0	0	0	16
5	0	0	0	0	35

I first default to everyone shedding, and then gradually change to not shedding. Data were not missing (i.e., panss scores were still available until 18 months, meaning follow-up was completed) and were considered right-censored. Details are shown in Table 22. For example, when `panss9` is missing and `panss3` is available, it is regarded as falling off after 3 months (time point is 12).

Table 22: Definition of Dropout Time Based on Missingness Pattern

Missingness Pattern	**Assumed Dropout Time (weeks)**
`panss18` observed	0 (Not dropped out)
`panss18` missing, `panss9` observed	36
`panss9` missing, `panss3` observed	12
`panss3` missing, `panss1` observed	6
`panss1` missing, `panss0` observed	6

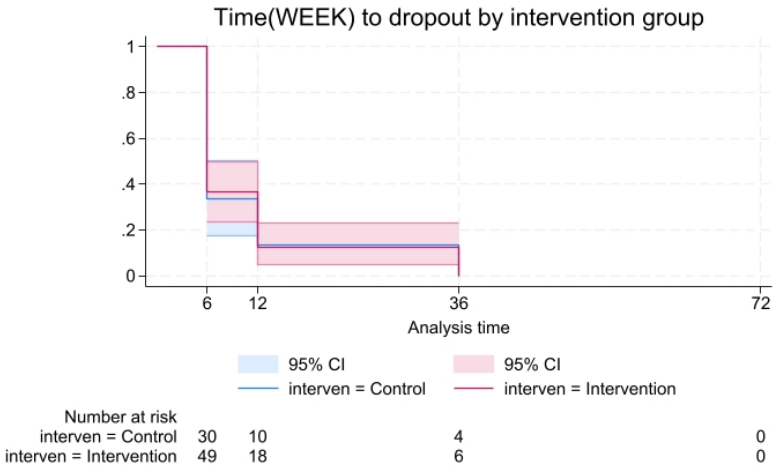

Figure 8: Kaplan-Meier survival curve

Intuitively, the survival probability of the two groups changed almost the same over time, although there was a certain tendency to separate at 3 months. The survival rate stabilized after 12 weeks, indicating that withdrawal was concentrated in the early stage, which is similar to the results of the previous analysis.

Table 23: Log-rank Test for Equality of Survivor Functions by Group

Group	**Observed Events**	**Expected Events**
Control	30	29.80
Intervention	49	49.20
Total	79	79.00
Chi-square (χ^2)(1) = 0.01, $p = 0.9326$		

There was no significant difference between the survival function (i.e., exit rate curve) of the intervention group and the control group (p = 0.9326), indicating that the exit patterns of the two groups were basically the same.

Section 11 Joint Model

In the process of building the joint model, after I included the interaction term, the interaction term was not statistically significant. Again, I chose to delete the interaction item. Then I fit a joint model with other covariates. Patients with poor or rapidly declining

mental health are at increased risk of dropping out due to inadequate response (Henderson et al., 2000).This indicates that a potentially influential variable in the model, the higher the PANSS score, the greater the dropout risk of patients

Table 24: Joint Model without Interaction

Parameter	**Coef.**	**Std. Err.**	**z**	**p-value**	**95% CI**
Longitudinal Submodel					
trt2 (Intervention)	-3.86	1.90	-2.03	0.042	[-7.58, -0.14]
logDUP	7.29	1.49	4.90	0.000	[4.38, 10.21]
Baseline PANSS	0.48	0.05	9.15	0.000	[0.38, 0.59]
Intercept	14.79	5.09	2.90	0.004	[4.81, 24.78]
Survival Submodel					
Association (Intercept)	0.043	0.021	2.07	0.038	[0.002, 0.084]
trt2 (log hazard)	-0.68	0.32	-2.15	0.032	[-1.29, -0.06]
Intercept (ln lambda)	-4.19	0.63	-6.69	0.000	[-5.41, -2.96]
ln gamma	-0.47	0.14	-3.29	0.001	[-0.75, -0.19]
Random Effects					
sd(Intercept)	9.54	0.74	–	–	[8.18, 11.11]
sd(Residual)	11.60	0.38	–	–	[10.87, 12.37]

From Table 24, in the Longitudinal part, Coefficient of intervention group compared with control group was -3.86 (95% CI: -7.58,-0.14), which I could conclude that the intervention group had significantly lower PANSS scores, meaning greater improvement in symptoms after controlling other variable. This was the same finding with that in the Session4. While in the Survival Submodel, the association parameter represents the sensitivity of the survival submodel to the current value of PANSS. And its was 0.043, P-Value is less 0.05,which was statistically significant. Indicating that when PANSS score was higher, patient's probability of drop-out would be higher. Also, the estimated of treatment effect in survival submodel was -0.68, illustrating that the withdrawal was more pronounced in the intervention group than in the control group.

Compare it with model in the Session 4, and to make the comparison more intuitive, I

fit another model with no logarithmic transformation of the dependent variable.

Table 25: Mixed Effects Model Results for PANSS Score

Variable	**Coef.**	**Std. Err.**	**z**	**p-value**
Fixed Effects				
Month (change)	-0.215	0.071	-3.04	0.002
Intervention	-4.704	1.632	-2.88	0.004
Baseline PANSS	0.479	0.047	10.28	¡0.001
log(DUP)	7.738	1.336	5.79	¡0.001
Intercept	14.031	4.499	3.12	0.002
Random Effects (Unstructured)				
Var(Month slope)	0.371	0.123		
Var(Intercept)	140.130	24.969		
Cov(Month, Intercept)	3.301	1.475		
Residual Variance	125.502	9.120		

Comparing the two models, as well as the model 1 established before session 4, both models showed a decrease in PANSS score under the intervention. In terms of fixed effects, the estimated coefficient of the intervention group in GLMM was -4.70. The estimated coefficient of the combined model is -3.86, which is slightly lower than that of the combined model. However, the direction of intervention was consistent and statistically significant. In my opinion, participants' withdrawal would not materially affect the estimate of treatment effect, or this would not affect the results directly. And this could also indicate the robustness of the results.

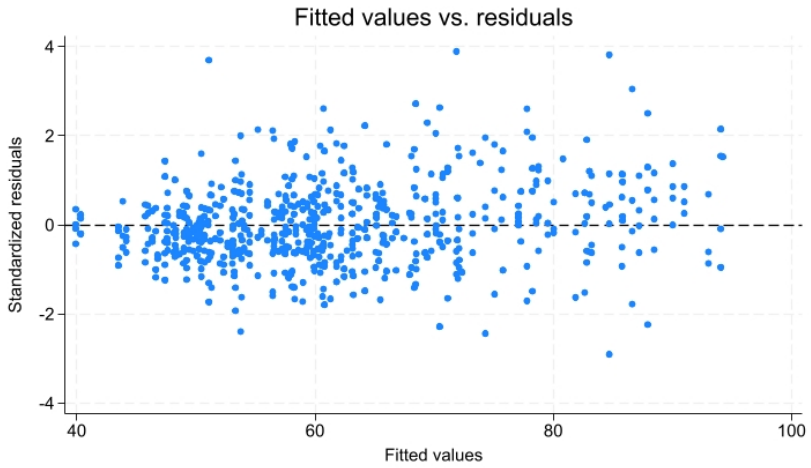

Figure 9: Residual in the Joint Model

From the scatterplot of the standardized residuals and fitted values of the joint model, I found that the residuals were roughly symmetric around 0, which showed that there was no significant systematic bias. In my view this model was fitted well. At the same time, there was no curve trend, the assumption of linearity was satisfied, and there was no heteroscedasticity.

Reference

Henderson, R. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, *1*(4), 465–480. https://doi.org/10.1093/biostatistics/1.4.465

```
//Remove duplicate and unpronounced double-key words
#include <bits/stdc++.h>
using namespace std;
int main(){
    freopen("input.txt","r",stdin);
    freopen("output.txt","w",stdout);
    int n;
    cin >> n;
    string s[n];
    for(int i=0;i<n;i++){
        cin >> s[i];
    }
    int count=0;
    for(int i=0;i<n;i++){
        if(s[i]==s[i+1]){
            count++;
        }
    }
    cout << count << endl;
    for(int i=0;i<n;i++){
        if(s[i]==s[i+1]){
            s[i]="";
        }
    }
    for(int i=0;i<n;i++){
        cout << s[i] << " ";
    }
    return 0;
}
```

```
#include<bits/stdc++.h>
using namespace std;
int main()
{
    int n;
    cin>>n;
    int a[n];
    for(int i=0;i<n;i++)
    {
        cin>>a[i];
    }
    int count=0;
    for(int i=0;i<n;i++)
    {
        if(a[i]==a[i+1])
        {
            count++;
        }
    }
    cout<<count<<endl;
    for(int i=0;i<n;i++)
    {
        if(a[i]==a[i+1])
        {
            a[i]=0;
        }
    }
    for(int i=0;i<n;i++)
    {
        if(a[i]!=0)
        {
            cout<<a[i]<<" ";
        }
```

```
//7PAVMALM Multilevel and Longitudinal Modelling module
// name
//Coursework
clear all


cd "C:\Users\17789\Desktop\Multilevel and Longitudinal Modelling\coursework"
use "Socrates_assignment dataset.dta"
//log using "assignment.log", replace

label list // I found that I needed to solve some data
tab substmis
//tab2 centre idnumber
/*
DATA Cleaning
*/
gen male=2-sex // binary variabl, so female is "0"; male is "1"
//replace substmis = 0 if substmis == 8
//replace substmis = . if substmis == 7
label variable male "sex recoded to binary"

/*Question_1*/
misstable summarize panss0 panss1 panss3 panss9 panss18
mdesc panss0 panss1 panss3 panss9 panss18
misstable pattern panss0 panss1 panss3 panss9 panss18, asis
misstable pattern panss0 panss1 panss3 panss9 panss18, freq bypattern
misstable tree panss0 panss1 panss3 panss9 panss18, asis

//xtdescribe if panss<.

//gen missing_panss19 = missing(panss9)
//logit missing_panss19 interven sex ageentr yearsofe substmis panss0

/*Question_2*/
tabstat panss0remis panss1remis panss3remis panss9remis panss18remis, by(interven) statistics(mean) format(%9.2f) //Response rates were counted by group and time point

*But I want to check whether the above answer is correct, and I choose to convert the wide data to the long data for testing
rename panss0remis panssremis0
rename panss1remis panssremis1
rename panss3remis panssremis3
rename panss9remis panssremis9
rename panss18remis panssremis18
```

```
*Converts wide data to long data
frame copy default default_cp
frame rename default wide_data
*******Reshaping
frame default_cp{
            reshape long panss panssremis, i(idnumber) j(time)
}

frame rename default_cp long_data

frame change long_data

replace time = 1.5 if time == 1
* I don't want to change the original "reshaped" data. Keep it
preserve


*Average response rates were calculated by time and intervention group
collapse (mean) panssremis, by(time interven)
summarize panssremis // check the answer
//tabout panssremis using "output.tex", sum format(%9.2f) replace style(tex) label

*graph 1
twoway ///
      (line panssremis time if interven == 0, lcolor(edkblue) lpattern(dash)) ///
      (line panssremis time if interven == 1, lcolor(eltgreen) lpattern(solid)), ///
      xlabel(0 "Base" 1.5 "6 Weeks" 3 "3 Months" 9 "9 Months" 18 "18 Months", angle(45)) ///
      ytitle("Remission Rate(%)") ///
      xtitle("Time") ///
      legend(label(1 "Routine Care Arm") label(2 "Intervention Arm")) ///
      title("PANSS Remission Rate Over Time by Group")



*Restore to the "reshaped" data
restore


/*Question_3*/
preserve
*Average PANSS scores were calculated by time and intervention group
collapse (mean) panssmean = panss (sd) pansssd = panss pansscount = panss, by(time interven)
```

```
gen upper=panssmean+1.96*(1/sqrt(pansscount))*pansssd
gen lower=panssmean-1.96*(1/sqrt(pansscount))*pansssd
// Jittering the X-axis time points to stagger the data points slightly
gen time1=time-0.02
gen time2=time+0.02
//summarize panss
*graph 2
******line gragh
twoway (line panssmean time1 if interven == 0, lcolor(edkblue) lpattern(dash)) ///
        (line panssmean time2 if interven == 1, lcolor(eltgreen) lpattern(solid)) ///
        (rcap lower upper time1 if interven==0, blcolor(edkblue)) ///
        (rcap lower upper time2 if interven==1, blcolor(eltgreen)), ///
        xlabel(0 "Base" 1.5 "6 Weeks" 3 "3 Months" 9 "9 Months" 18 "18 Months", angle(45)) ///
        ytitle("Mean PANSS Score") xtitle("Time") ///
        legend(order(1 "Routine Care Arm" 2 "Intervention Arm" 3 "95% CI" 4 "95% CI") ///
                col(2) position(6) ring(0.5)) ///
        title("Mean PANSS Score Over Time by Group")

* Recover "reshaped" data
restore
*graph 3
******Box plot
graph box panss, ///
    over(interven, label(labcolor(black) angle(45))) ///
    over(time, relabel(1 "Base" 2 "6 Weeks" 3 "3 Months" 4 "9 Months" 5 "18 Months")) ///
    asyvars ///
    box(1, fcolor(edkblue) lcolor(edkblue) ) ///    // Sets the fill, border, and dot color and style
for interven == 0
    box(2, fcolor(eltgreen) lcolor(eltgreen)) ///
    marker(1, mcolor(edkblue)) ///
    marker(2, mcolor(eltgreen)) ///
    title("PANSS Score Distribution Across Interventions and Time")

******Every pacients Line graph
sort idnumber time
twoway (line panss time, connect(ascending)), by(interven)


/*Question_4*/
frame change wide_data // chaging to wide data
frame copy wide_data default_cp_another //
frame change default_cp_another
rename panss0 base
rename panssremis0 baseremin
```

```
*******Reshaping
frame default_cp_another{
            reshape long panss panssremis, i(idnumber) j(time)
}

frame rename default_cp_another long_data_another
gen month=time
//replace month = 1.5 if month == 1

frame long_data_another {
****** TRANSFORM OUTCOME
      //boxcox panss, model(lhsonly)
      //generate BCpanss = (panss^(-0.1049326)-1)/(-0.1049326)
      //generate BCpanss_base = (base^(-0.1049326)-1)/(-0.1049326)
      //summarize BCpanss
      //mean BCpanss

      gen logpanss=log10(panss)
      label variable logpanss "PANSS score (log10)"
      // gen transformed=sqrt(panss-30) if panss>25
      // scatter logp transformed || function y = x, ra(logp) clpat(dash)

      gen logbase=log10(base+1)
      label variable logbase "PANSS score at baseline (log10 p+1)"
      gen month_change = month - 18
      gen month_cen = month - 9
      //gen panss_adjust = base -panss
}


//histogram BCpanss, normal kdensity

histogram logpanss, normal kdensity
frame change wide_data
frame wide_data {
          preserve
          gen logpanss0 = log10(panss0)
          gen logpanss1 = log10(panss1)
          gen logpanss3 = log10(panss3)
          gen logpanss9 = log10(panss9)
          gen logpanss18 = log10(panss18)
          corr logpanss1 logpanss3 logpanss9 logpanss18
          restore
}
```

```
frame change long_data_another
//xtdescribe if panss<. // A check of the first question
histogram panss, normal kdensity
*****null model
//mixed panss month_change ||idnumber:
mixed logpanss month_change || idnumber:
estat icc
estimates store model1_null
mixed logpanss i.interven c.month_change logbase c.logdup i.male i.substmis c.yearsofe i.episode c.ageentr || idnumber:
estat icc
estimates store model1_other
//lrtest (model1_null) (model1_other), stats
mixed logpanss c.month_change logbase i.interven c.logdup || idnumber:
estimates store model1_other1
estat ic
mixed logpanss c.month_change i.interven logbase c.logdup || idnumber: month_change, cov(unstr) mle
estimates store model1
lrtest (model1_other1) (model1), stats
estat ic
estat icc
mixed logpanss c.month_change##i.interven logbase c.logdup || idnumber: month_change, cov(unstr) mle
estimates store model12
lrtest (model12) (model1), stats
mixed panss c.month_change##i.interven base c.logdup || idnumber: month_change, cov(unstr) mle

********Assumption
estimates restore model1
predict residual, residual
graph box residual
qnorm residual, title("GLMM(Continued) Normal Q-Q Plot")
//predict reintercept, reffects
//graph box reintercept if month == 0

graph box residual, by(month_change)

********fingdings
margins interven, at(month = 0)
marginsplot, xline(0)
//quietly margins i.month_change, over(interven) predict(xb)
```

```
/*Question_5*/
iis idnumber
melogit panssremis c.month_cen i.interven baseremin c.logdup || idnumber:,or
estimates store model2_other


generate trt_month = interven*month_cen
melogit panssremis c.month_cen trt_month i.interven baseremin c.logdup || idnumber: month_cen, or

estimates store model2_other1

melogit panssremis c.month_cen i.interven baseremin c.logdup || idnumber: month_cen, or
estimates store model2
/*
melogit panssremis c.month_cen i.interven baseremin c.logdup || idnumber: month_cen, cov(unstr) or
estimates store model21
*/
lrtest (model2_other1) (model2), stats
lrtest (model2_other) (model2), stats

********Assumption
//predict remodel2, deviance conditional(fixedonly)    // incorrect
//qnorm remodel2, title("GLMM(Continued) Normal Q-Q Plot")
predict reffects*, reffects
qnorm reffects1
qnorm reffects2
describe reffects* // Comfirm the name
twoway scatter reffects2 reffects1, ///
       xlabel(-4(1)4) ylabel(-1(0.5)1) ///
       title("Joint distribution of random effects") ///
       xtitle("Random Intercept") ytitle("Random Slope")

twoway (scatter reffects2 reffects1) ///
           (lfitci reffects2 reffects1, level(90)), ///
           title("Random effects with 90% CI Ellipse")

corr reffects1 reffects2
/*
                    | reffec~1 reffec~2
-------------+------------------
    reffects1 |      1.0000
    reffects2 |    -0.0262      1.0000
```

```
*/

******findings

margins interven, at(month_cen = (0))    // 9 months
marginsplot
lincom 1.interven, or

/*Question_6*/
iis(idnumber)
gen month_ooo = month
replace month_oo = 6 if month_o == 1
replace month_oo = 12 if month_o == 3
replace month_oo = 36 if month_o == 9
replace month_oo = 72 if month_o == 18
xtset idnumber month_oo
/*
** 1. GEE with exchangeable correlation structure
xtgee panss logdup base i.interven##c.month_o, family(gaussian) link(identity) corr(exchangeable)
xtgee panss logdup base i.interven##c.month_o, family(gaussian) link(identity) corr(exchangeable) vce(robust)
estat wcor

** 2. GEE with Unstructured correlation structure
xtgee panss logdup base i.interven##c.month_o, family(gaussian) link(identity) corr(unstr)
xtgee panss logdup base i.interven##c.month_o, family(gaussian) link(identity)corr(unstr) vce(robust)
estat wcor
*/
* Exchangeable
gen interven_month = interven * month_oo
qic panss logdup base interven month_oo interven_month, ///
      i(idnumber) family(gaussian) link(identity) ///
      corr(exchangeable) robust

* Unstructured
qic panss logdup base interven month_oo interven_month, ///
      i(idnumber) family(gaussian) link(identity) ///
      corr(unstructured) robust

//From the QIC, I    choosed "EXCH"
xtgee panss logdup base i.interven##c.month_oo, family(gaussian) link(identity) corr(exchangeable)
```

```
xtgee panss logdup base i.interven##c.month_oo, family(gaussian) link(identity)corr(exchangeable) vce(robust)
estat wcor
xtgee panss logdup base i.interven month_oo, family(gaussian) link(identity) corr(exchangeable)
xtgee panss logdup base i.interven month_oo, family(gaussian) link(identity) corr(exchangeable) vce(robust)
estimates store model3
estat wcor
******Sensetive test
xtgee logpanss logdup logbase i.interven month_oo, family(gaussian) link(identity) corr(exchangeable) vce(robust)

xtgee panss logdup base i.interven month_change, family(gaussian) link(identity) corr(exchangeable) vce(robust) // Delecte the interaction



*********Assumption
//MAR MCAR MISSING
// GAUSSIAN IDENTITY
// Patients are individually related and independent of each other
vif, uncentered
/*Question_7*/
**********QIC
qic panssremis logdup baseremin interven month_oo, ///
      i(idnumber) family(bin) link(logit) ///
      corr(unstructured) robust

qic panssremis logdup baseremin interven month_oo, ///
      i(idnumber) family(bin) link(logit) ///
      corr(exchangeable) robust
**********Modelling
xtgee panssremis c.month_oo i.interven base c.logdup , family(binomial) link(logit) corr(exchangeable) eform

xtgee panssremis c.month_oo i.interven base c.logdup , family(binomial) link(logit) corr(exchangeable) eform vce(robust)
estat wcor

xtgee panssremis c.month_oo i.interven base c.logdup , family(binomial) link(logit) corr(unstructured) eform vce(robust)
estat wcor

xtgee panssremis c.month_cen i.interven base c.logdup , family(binomial) link(logit) corr(exchangeable) eform vce(robust)
```

```
xtgee panssremis c.month_oo##i.interven base c.logdup , family(binomial) link(logit) corr(exchangeable) vce(robust) eform
/*
xtgee panssremis, ///
	family(binomial) link(logit) corr(exchangeable) vce(robust)
*/

xtgee panssremis c.month_oo i.interven baseremin c.logdup, ///
	family(binomial) link(logit) corr(exchangeable) vce(robust)
	// interaction
estimates store model4
//vif
**********Assumption
vif, uncentered
/*Question_9*/

table centre interven, missing // centre and interven
// The center may still have a clustering effect
table interven therapis, miss // There is significant confusion between intervention and therapist

mixed logpanss c.month i.interven logbase c.logdup ///
	|| centre: , ///
	|| idnumber: month_change, cov(unstructured)


/*Question_10*/
frame change wide_data
frame copy wide_data monotonic_missing
frame change monotonic_missing

frame monotonic_missing{
gen monotonic_missing = 0

replace monotonic_missing = 1 if ///
	(missing(panss1) & missing(panss3) & missing(panss9) & missing(panss18)) | ///
	(!missing(panss1) & !missing(panss3) & !missing(panss9) & !missing(panss18)) | ///
	(!missing(panss1) & !missing(panss3) & !missing(panss9) & missing(panss18)) | ///
	(!missing(panss1) & !missing(panss3) & missing(panss9) & missing(panss18)) | ///
	(!missing(panss1) & missing(panss3) & missing(panss9) & missing(panss18))
keep if monotonic_missing == 1
}

misstable pattern panss1 panss3 panss9 panss18, asis
```

```
misstable pattern    panss1 panss3 panss9 panss18, freq bypattern

gen dropout_time = .
gen dropout_event = 1      // Default as off

// People who completed follow-up: With data available at 18 months, marked as right-censored
replace dropout_time = 0 if !missing(panss18)
replace dropout_event = 0 if dropout_time == 0

// Off from 9 months
replace dropout_time = 36 if missing(panss18) & !missing(panss9)& !missing(panss3)
& !missing(panss1)

// Off from 3 months
replace dropout_time = 12 if missing(panss9) & !missing(panss3)

// Off from 6 weeks
replace dropout_time = 6 if missing(panss3) & !missing(panss1)

// Shedding immediately after baseline (panss1 loss = no 6 weeks)
replace dropout_time = 6 if missing(panss1) & !missing(panss0)

// save dropoutinfo, replace

// Complete follow-up is considered "not left behind"
stset dropout_time, failure(dropout_event = 1) id(idnumber)
sts graph, by(interven) ci ///
      title("Time(WEEK) to dropout by intervention group") ///
      xlabel(6 12 36 72) ///
      ylabel(0(.2)1) ///
      risktable ///
      legend(position(6) ring(0.5))

sts test interven, logrank
sts test interven, wilcoxon
sts test interven, tware
sts graph, hazard


/*Question_11*/
frame copy monotonic_missing joint
frame change joint

frame copy monotonic_missing joint2
```

```
frame change joint2
frame copy monotonic_missing joint3
frame change joint3
frame copy monotonic_missing joint4
frame change joint4
frame copy monotonic_missing joint5
frame change joint5
frame copy monotonic_missing joint6
frame change joint6
frame copy monotonic_missing joint7
frame change joint7
frame copy monotonic_missing joint8
frame change joint8
frame copy monotonic_missing joint9
frame change joint9

egen nobs=rownonmiss(panss*)
rename panss0 base
rename panss0remis panssremis0
rename panss1remis panssremis1
rename panss3remis panssremis3
rename panss9remis panssremis9
rename panss18remis panssremis18
//stset survtime ,failure(event==1) id(idnumber)
reshape long panss panssremis, i(idnumber) j(time)
recode time (0=0) (1=6) (3=12) (9=36) (18=72), generate(week)
drop if panss==.
sort idnumber week
bysort idnumber: egen survtime = max(week)
// Prepare survival data for the start-stop structure
bysort idnumber: gen start = week
bysort idnumber: gen stop = start[_n+1]
//Initialize the event indicator variable (1 = dropout, 0 = censored)
gen event = 0
//gen episode1st = 2-episode
bysort idnumber: replace stop = survtime if _n==_N
bysort idnumber: replace event = dropout_event if _n==_N
replace stop = start + 0.01 if stop <= start
//drop if stop <= start
bysort idnumber (week): replace stop = survtime if _n == _N

tab interven, gen(trt)
tab interven
gen duration = stop - start
```

```
list idnumber start stop if duration <= 0

stset stop, enter(start) failure(event==1) id(idnumber)

/*
stjm panss trt2, ///
      panel(idnumber) survmodel(weibull) ffp(1) ///
      intassociation nocurrent
*/


stjm panss trt2 logdup base, ///
      panel(idnumber) survmodel(weibull) ffp(1) ///
      timeinteraction(trt2) ///
      survcov(trt2 logdup base) ///
      intassociation nocurrent gh(10)    // interaction


stjm panss trt2 logdup base, ///
      panel(idnumber) survmodel(weibull) ffp(1) ///
      survcov(trt2 logdup base) ///
      intassociation nocurrent gh(10)

estimates table
//Residuals and fitted values
predict fittedvals1, fitted longitudinal
predict resids1, rstandard
scatter resids1 fittedvals1, yline(0) ytitle("Standardized residuals") xtitle("Fitted values")
title("Fitted values vs. residuals")

/* Predicted survival */
predict survfit, xb survival
sts gen km=s km_lci=lb(s) km_uci=ub(s), by(interven)

twoway (rarea km_lci km_uci _t if interven==1,sort col(gray*0.6)) ///
(line km _t if interven==1,sort)(line survfit _t if interven==1,sort lpat(solid)) ///
, plotr(m(zero)) ylabel(0(0.2)1, angle(h) format(%2.1f))       ///
ytitle("Survival probability") xtitle("Follow-up (months)")    ///
title("Predicted marginal survival") legend(order(1 "95% Confidence Interval" ))
label list
//stjmcsurv, panel(id) id(2) fu(15) save(g1, replace)
//stjmcsurv, panel(id) id(102) fu(15) save(g2, replace)
//graph combine (g1 g2)
nlcom [alpha_1][_cons]*[Longitudinal][trt] + [ln_lambda][trt]
```

```
/*
predict xb_long, xb
// Predicted values of longitudinal submodels (PANSS)
predict survprob, survival
// Predicted survival probability

stjm panss trt1 trt2, ///
      panel(idnumber) survmodel(weibull) ffp(1) ///
      timeinteraction(trt1 trt2) ///
      survcov(trt1 trt2) ///
      intassociation nocurrent gh(10)

stjm panss trt1 trt2, ///
      panel(idnumber) survmodel(weibull) ffp(1) ///
      survcov(trt1 trt2) ///
      intassociation nocoef
estimates table


streg i.interven, distribution(weibull) nohr
stcurve, surv at1(treat=1) at2(treat=2) at3(treat=3)
*/
/*
gen month = time
recode month (1=6) (3=12) (9=36) (18=72), generate(week)
drop month week
reshape wide panss panssremis, i(idnumber) j(time)

stset survtime, failure(inform=1) id(idnumber)
streg i.treat, distribution(weibull) nohr
stcurve, surv at1(treat=1) at2(treat=2) at3(treat=3)


// Prepare survival data for the start-stop structure
bys idnumber (week): gen start = week
bys idnumber (week): gen stop = start[_n+1]

bys idnumber (week): gen survtime = week if _n==_N
//* Initialize the event indicator variable (1 = dropout, 0 = censored)

//Initialize the event indicator variable (1 = dropout, 0 = censored)
gen event1 = 0
```

```
bys idnumber: replace stop = survtime if _n==_N
bys idnumber: replace event1 = dropout_event if _n==_N

tab interven, gen(trt)
stset stop, enter(start) failure(event1==1) id(idnumber)
stsum
stjm panss trt, panel(idnumber) ///
        survmodel(weibull) ffp(1) ///
        survcov(trt) ///
        timeinteraction(trt) ///
        intassociation gh(10)
estimates table

predict resid, rstandard
scatter resid _t
*/
```