Longitudinal Analysis for final project

Shih-Ni Prim

11/12/2020

Contents

To see whether Drug 1 and Drug 2 had different effects, we performed a hypothesis test on $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ by testing

$$\mathbf{L_4}\beta = 0$$

where

$$\mathbf{L_4} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_10, \beta_11)^T$

1 Data Analysis

1.1 Longitudinal Data Analysis

To answer our research question Q2, **How does the binding strength of the antibodies develop in response to the number of vaccine dosages by treatment?**, we use longitudinal data analysis, including general linear models and linear mixed models.

As seen in Figure 1 and Figure 2, the mean trend is not linear, and the different time points have different variances. This information suggests that we should use piecewise linear models and set variances as unequal over time.

[Figure 1 about here.]

[Figure 2 about here.]

We first consider a model with time point as the only covariate:

$$Y_{ij} = \beta_0 + \beta_1 Time_{ij} + e_{ij}$$

Thus we will use a piecewise linear model, in which each segment has different intercepts and slopes. We use three indicator variables: $S1, S2, S3$ as the indicator variables, where

$$S1 = \begin{cases} 1 & \text{if } 0 \leq \text{Timepoint} < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$S2 = \begin{cases} 1 & \text{if } 1 \leq \text{Timepoint} < 2 \\ 0 & \text{otherwise} \end{cases}$$

$$S3 = \begin{cases} 1 & \text{if Timepoint} \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

The new model is thus

$$Y_{ij} = S1(\beta_0 + \beta_1 Time_{ij}) + S2(\beta_2 + \beta_3 Time_{ij}) + S3(\beta_4 + \beta_5 Time_{ij}) + e_{ij}$$

We also want to make sure that the trend is continuous at timepoint = 1 and 2.

Our model is $Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) + \beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + e_{ij}$ where

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

Again, our final mean model is

$$Y_{ij} = \beta_0(S1 + 2S2 - S2Time_{ij}) + \beta_1(S1Time_{ij} + 2S2 - S2Time_{ij}) +$$

$$\beta_4(-S2 + S2Time_{ij} + S3) + \beta_5(-2S2 + 2S2Time_{ij} + S3Time_{ij}) + e_{ij}$$

which can be written as

$$Y_{ij} = S1(\beta_0) + S1Time_{ij}(\beta_1) + S2(2\beta_0 + 2\beta_1 - \beta_4 - 2\beta_5) + S2Time_{ij}(-\beta_0 - \beta_1 + \beta_4 + 2\beta_5)$$

$$+S3(\beta_4) + S3Time_{ij}(\beta_5) + e_{ij}$$

From the model above, we can find the intercepts and slopes for all three segments of the mean trend and make a plot, as seen in Figure 3:

S1: $-0.2221651 + 0.2432183 * time$

S2: $(2 * -0.2221651 + 2 * 0.2432183 - 0.7699600 + 2 * 0.2432756) + (0.2221651 - 0.2432183 + 0.7699600 - 2 * 0.2432756) * time = -0.2413024 + 0.2623556 * time$

S3: $0.7699600 - 0.2432756 * time$

[Figure 3 about here.]

As shown in Figure 3, the two segments S1 and S2 look linear. So now we'll refit the model with only two piecewise sections; we'll call them S4 and S5. The new model is therefore

$$Y_{ij} = S4(\beta_0 + \beta_1 Time_{ij}) + S5(\beta_2 + \beta_3 Time_{ij}) + e_{ij}$$

$$S4 = \begin{cases} 1 & \text{if Timepoint} < 2 \\ 0 & \text{otherwise} \end{cases}$$

$$S5 = \begin{cases} 1 & \text{if Timepoint} \geq 2 \\ 0 & \text{otherwise} \end{cases}$$

We also want to make sure that the trend is continuous at Time_Point = 2.

Our model is then $Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + e_{ij}$ where

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

Again, our model is $Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + e_{ij}$, which can be written as $Y_{ij} = S4(-2\beta_1 + \beta_2 + 2\beta_3) + S4Time_{ij}(\beta_1) + S5(\beta_2) + S5Time_{ij}(\beta_3) + e_{ij}$

We first find the mean trend for S4 and S5:

S4: $(-2*0.5310975 + 0.5720853 + 2*-0.0000723) + 0.5310975*time = -0.4902543 + 0.5310975*time$ S5: $0.5720853 - 0.0000723*time$

We can make the plot again to see if the model is reasonable, as shown in Figure 4. Indeed, there is a linear line between Time_Point 0 and 2 and one between Time_Point 2 and 3. The two lines are contiuous at Time_Point 2. A comparison of AIC And BIC of these two models, shown in Table 1, indicates that the second model (`fit.gls2`) is indeed a better model. We'll add random effects to it.

[Figure 4 about here.]

[Table 1 about here.]

Next we check whether adding random effects improve the model. We assume that random effects exist in the intercept and slope. Our linear mixed model is then: $Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) + b_{0i} + b1iTime_{ij} + e_{ij}$
where

$$\mathbf{b}_i \sim N\left(0, \mathbf{D} = \begin{pmatrix} D_{11} & D_{12} \\ & D_{22} \end{pmatrix}\right)$$

and

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

[Table 2 about here.]

As shown in Table 2, the model `fit.a2` (random intercept and slope, AR1 correlation, unequal variances) has the lowest AIC And BIC, so it seems the best model. We now check residuals for both models.

[Figure 5 about here.]

All of the Q-Q plots in Figure 5 are reasonable, so we'll use `fit.a2` for further analysis.

Now we would like to know if the slopes between Time_Point 0 and 2 and between Time_Point 2 and 3 equal zero. H_0 : slope of $S4 = 0$ and slope of $S5 = 0$, which means $H_0 : \beta_1 = 0$ and $\beta_3 = 0$

Thus, we can check for two tests:

$$\mathbf{L_1} = 0$$

where $\mathbf{L_1} = (1, 0, 0)$ and $= (\beta_1, \beta_2, \beta_3)^T$ and

$$\mathbf{L_2} = 0$$

where $\mathbf{L_2} = (0, 0, 1)$ and $= (\beta_1, \beta_2, \beta_3)^T$

[Table 3 about here.]

As shown in Table 3, the slop of S4 has a very small p-value, while the slope of S5 is quite large, indicating that the change in Binding rate between Time_Point 0 and Time_Point 2 is significant while the change between Time_Point 2 and Time_Point 3 is not significant. We conclude that Time_Point 2, when the monkeys had received two vaccines, had the highest Binding rate, while the last vaccine shot at Time_Point 3 did not make a difference to the Binding rate.

1.2 Add drugs as a covariate

Next we add drugs as a covariate to see if it has effects on Binding. We use two indicator variables: `D2` and `D3`, where

$$D2 = \begin{cases} 1 & \text{if Drug = 2} \\ 0 & \text{otherwise} \end{cases}$$

$$D3 = \begin{cases} 1 & \text{if Drug = 3} \\ 0 & \text{otherwise} \end{cases}$$

Assuming that the random effects are the same for each drug, our full model is:

$$Y_{ij} = \beta_1(-2S4 + S4Time_{ij}) + \beta_2(S4 + S5) + \beta_3(2S4 + S5Time_{ij}) +$$

$$\beta_4 D2(-2S4 + S4Time_{ij}) + \beta_5 D2(S4 + S5) + \beta_6 D2(2S4 + S5Time_{ij}) +$$

$$\beta_7 D3(-2S4 + S4Time_{ij}) + \beta_8 D3(S4 + S5) + \beta_9 D3(2S4 + S5Time_{ij}) + b_{0i} + b1iTime_{ij} + e_{ij}$$

where

$$\mathbf{b}_i \sim N\left(0, \mathbf{D} = \begin{pmatrix} D_{11} & D_{12} \\ & D_{22} \end{pmatrix}\right)$$

and

$$\mathbf{e}_i \sim N(0, \sigma^2 I)$$

```
## Linear mixed-effects model fit by REML
##  Data: dataLDA1
##        AIC      BIC    logLik
##   3207.673 3306.373 -1586.836
##
## Random effects:
##  Formula: ~time | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.6424735 (Intr)
## time        0.6319522 -0.999
## Residual    0.2329293
##
## Correlation Structure: AR(1)
##  Formula: ~1 | id
##  Parameter estimate(s):
##       Phi
## 0.4172276
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | time
##  Parameter estimates:
##         1         0         2         3
## 1.0000000 0.4112624 6.0679296 5.7289228
## Fixed effects: list(meanform3)
##         Value Std.Error   DF    t-value p-value
## v1  0.3490556 0.3191170 2438  1.0938170  0.2741
## v2 -1.1765071 0.5022272 2438 -2.3425795  0.0192
## v3  0.7940787 0.3682851 2438  2.1561521  0.0312
## v4  0.2970310 0.4663132 2438  0.6369775  0.5242
## v5 -1.4458003 0.7399858   17 -1.9538217  0.0674
## v6  0.8573187 0.5404821 2438  1.5862111  0.1128
## v7 -0.1785046 0.4859598 2438 -0.3673237  0.7134
## v8  0.3263319 0.7369987   17  0.4427849  0.6635
## v9 -0.2201034 0.5500331 2438 -0.4001638  0.6891
##  Correlation:
##    v1     v2     v3     v4     v5     v6     v7     v8
## v2 -0.610
## v3  0.846 -0.938
## v4 -0.684  0.418 -0.579
## v5  0.414 -0.679  0.637 -0.605
## v6 -0.577  0.639 -0.681  0.842 -0.938
## v7 -0.657  0.401 -0.556  0.449 -0.272  0.379
## v8  0.416 -0.681  0.639 -0.285  0.462 -0.436 -0.634
## v9 -0.567  0.628 -0.670  0.388 -0.426  0.456  0.863 -0.937
```

```
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -1.19647370 -0.31178413 -0.10810405  0.05433803 14.06000633
##
## Number of Observations: 2464
## Number of Groups: 20
```

Now we want to find whether the drugs have any effects. To see whether Drug 1 and Drug 2 have any difference, we want to perform a hypothesis test on $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$, thus we can do the test

$$\mathbf{L_3} = 0$$

where

$$\mathbf{L_3} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

To see whether Drug 1 and Drug 3 have any difference, we want to perform a hypothesis test on $H_0 : \beta_7 = \beta_8 = \beta_9 = 0$, thus we can do the test

$$\mathbf{L_4} = 0$$

where

$$\mathbf{L_4} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

To see whether Drug 2 and Drug 3 have any difference, we want to perform a hypothesis test on $H_0 : \beta_4 = \beta_7, \beta_5 = \beta_8, \beta_6 = \beta_9$, thus we can do the test

$$\mathbf{L_5} = 0$$

where

$$\mathbf{L_5} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

and $= (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)^T$

We found that, as shown in Table 4, Drug 1 and Drug 2 do not have significantly different effects on Binding rates. As shown in Table 5, Drug 1 and Drug 3 do not have significantly different effects on Binding rates. Also, as shown in Table 6, Drug 2 and Drug 3 do not have significantly different effects on Binding rates. In other words, drug groups do not have signifant effects on our longitudinal model. Thus we will retain `fit.a2` as our best model.

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

List of Figures

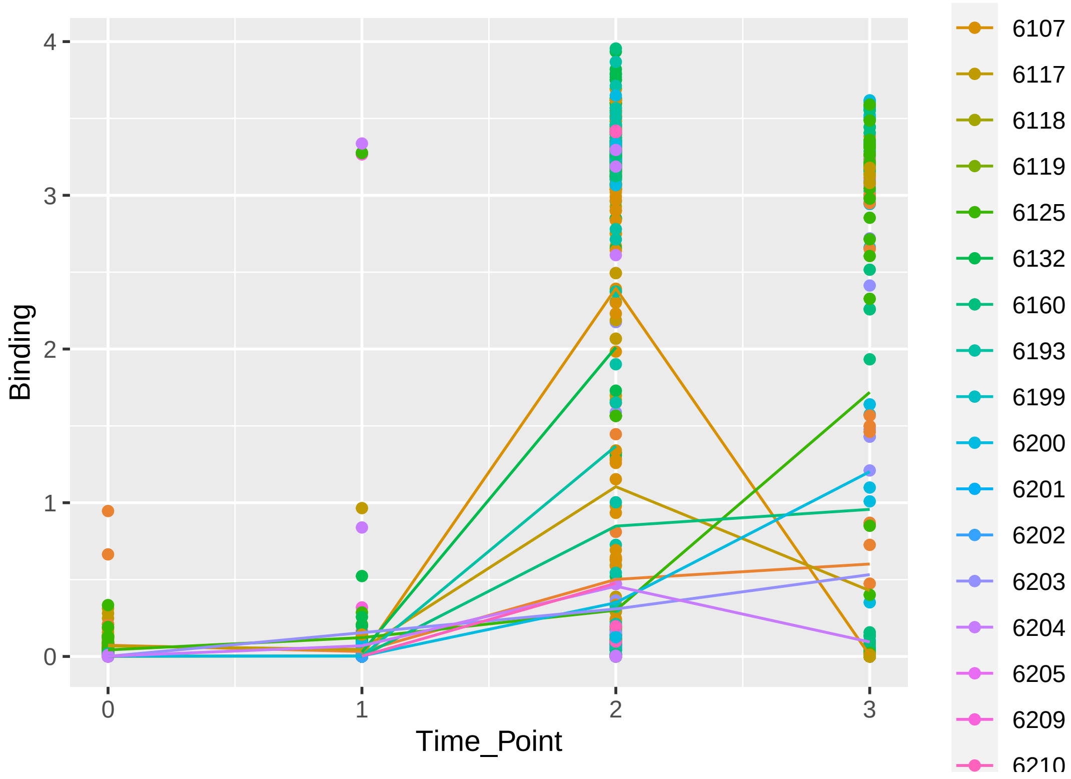

Figure 1: Mean trend by monkey

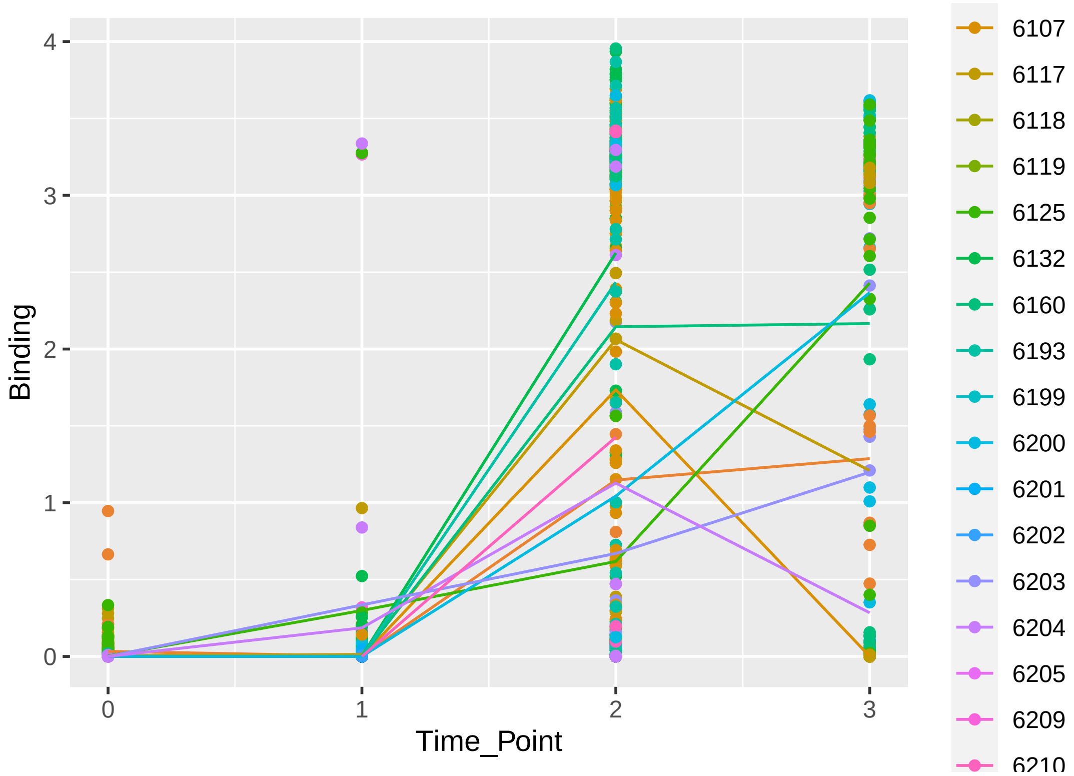

Figure 2: Variances over time by monkey

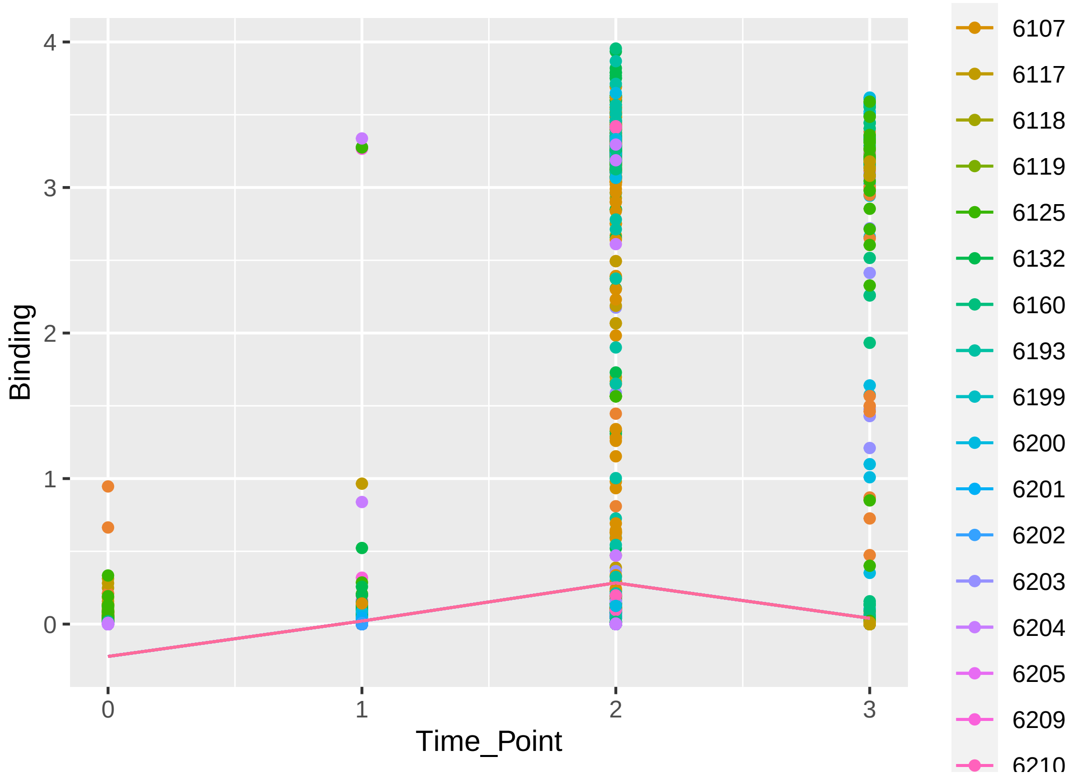

Figure 3: Piecewise Linear Function–three segments

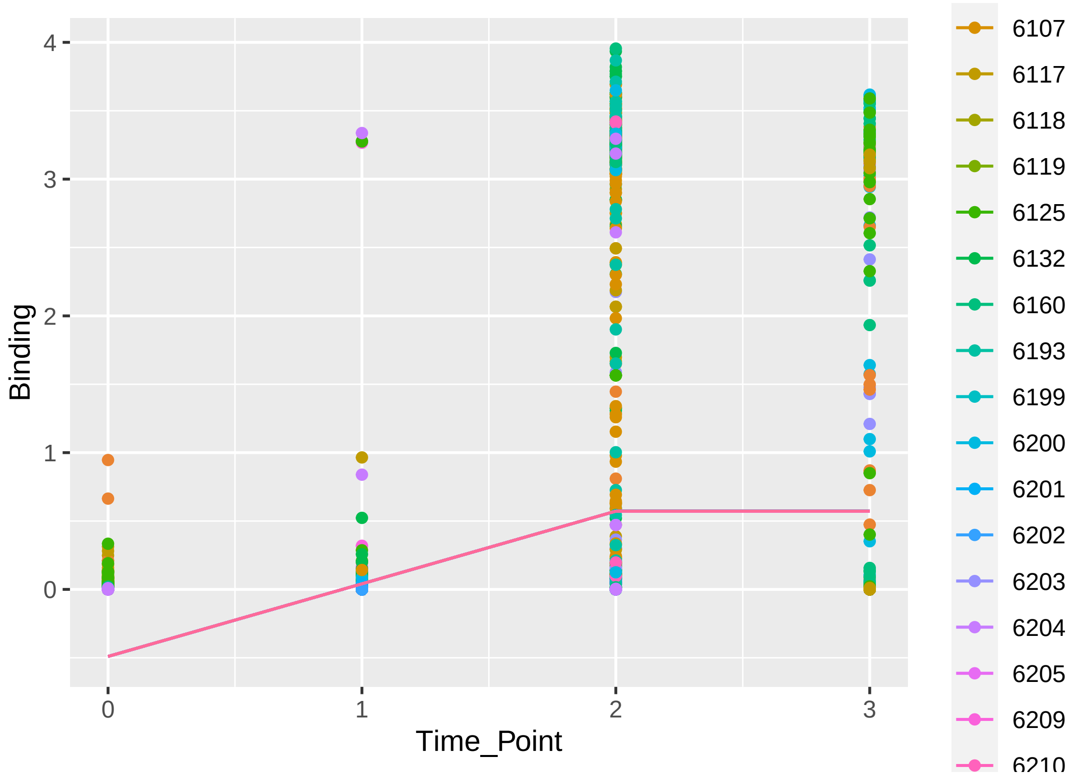

Figure 4: Piecewise Linear Function–two segments

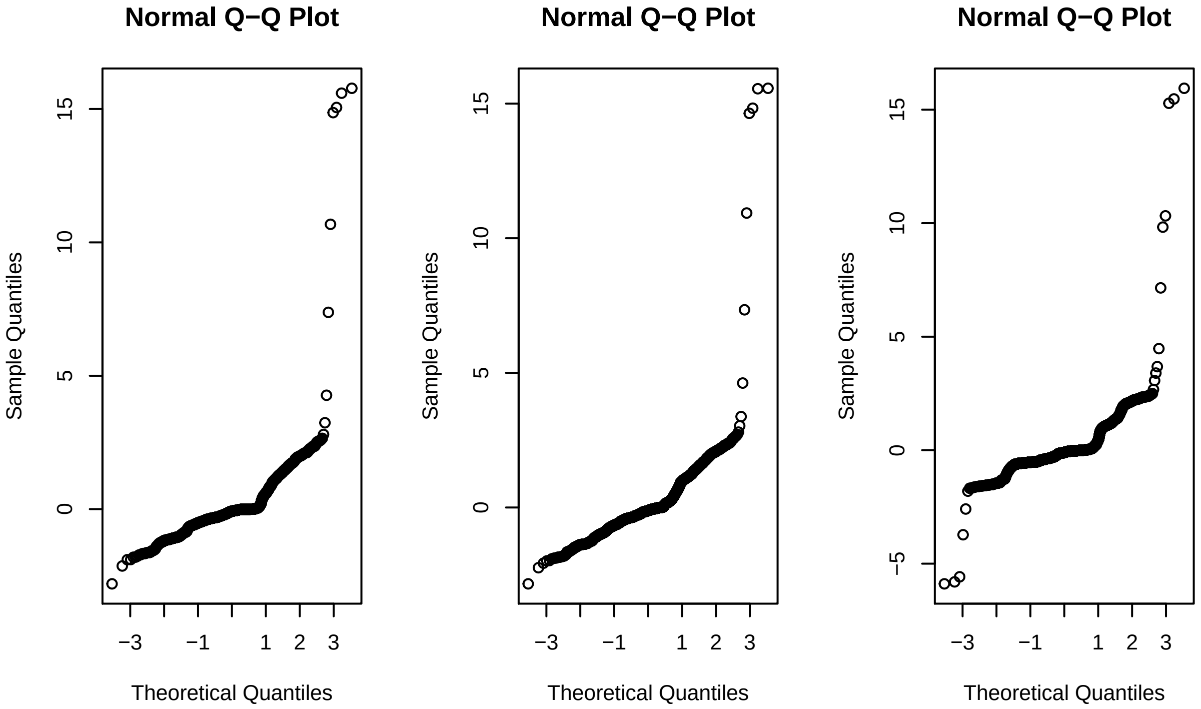

Figure 5: Q-Q plots of models: GLS, compound symmetry, AR1

List of Tables

Table 1: AIC and BIC between two gls models

	df	AIC	df.1	BIC
fit.gls	9	3323.050	9	3375.322
fit.gls2	8	3315.264	8	3361.730

Table 2: AIC and BIC for three models

	df	AIC	df.1	BIC
fit.gls2	8	3315.264	8	3361.730
fit.a1	11	3234.628	11	3298.520
fit.a2	11	3063.290	11	3127.182

Table 3: Inference of S4 ad S5 slopes

numDF	denDF	F.value	p.value
1	2442	244.2324506	0.0000000
1	2442	0.0317192	0.8586602

Table 4: Test whether drug 1 = drug 2

Fstat	p_value
1.065151	0.3626666

Table 5: Test whether drug 1 = drug 3

Fstat	p_value
1.231968	0.2964737

Table 6: Test whether drug 2 = drug 3

Fstat	p_value
1.255448	0.288075