Final Project – Second Draft

Kan Luo, Shih-Ni Prim

10/23/2020

Contents


```
## Response [https://raw.githubusercontent.com/luokan1227/537P1/master/Data.xlsx]
##   Date: 2020-10-24 05:01
##   Status: 200
##   Content-Type: application/octet-stream
##   Size: 341 kB
## <ON DISK>  C:\Users\shihn\AppData\Local\Temp\RtmpGqk185\file29a02eb76534.xlsx
```

```
## Response [https://raw.githubusercontent.com/luokan1227/537P1/master/MonkeyID.xlsx]
##   Date: 2020-10-24 05:01
##   Status: 200
##   Content-Type: application/octet-stream
##   Size: 50.1 kB
## <ON DISK>  C:\Users\shihn\AppData\Local\Temp\RtmpGqk185\file29a018ab1b18.xlsx
```

Introduction

Methodologies

Data Summaries

Contingency Tables

The study included 20 rhesus macaques.

```
table(Data2$MonkeyID)
```

```
##
## 6104 6105 6107 6117 6118 6119 6125 6132 6160 6193 6199 6200 6201 6202 6203 6204
##   35  228  239  243    7   55  216  251  183  117   48  191   73   78  238  156
```

```
## 6205 6209 6210 6214
##    5   46   50    6
```

There are four time points; one before any procedure was done, and three after vaccine shots were administered to the macaques. In the treatment groups, groups 1-3 represent different doses of drug 1, groups 4-6 represent different doses of drug 2, and group 7 represents the control group. Later we'll look at the effect made by different drugs first and then different doses.

```
table(Data2$Time_Point, Data2$Treatment)
```

```
##
##     group 1 group 2 group 3 group 4 group 5 group 6 group 7
##   0     129       0       0      90       0       0      54
##   1     190      60      96     105     297     131     125
##   2     141     110       0     148      77       0     347
##   3     122       0       0     101       0       0     142
```

```
table(Data2$Drug, Data2$Treatment)
```

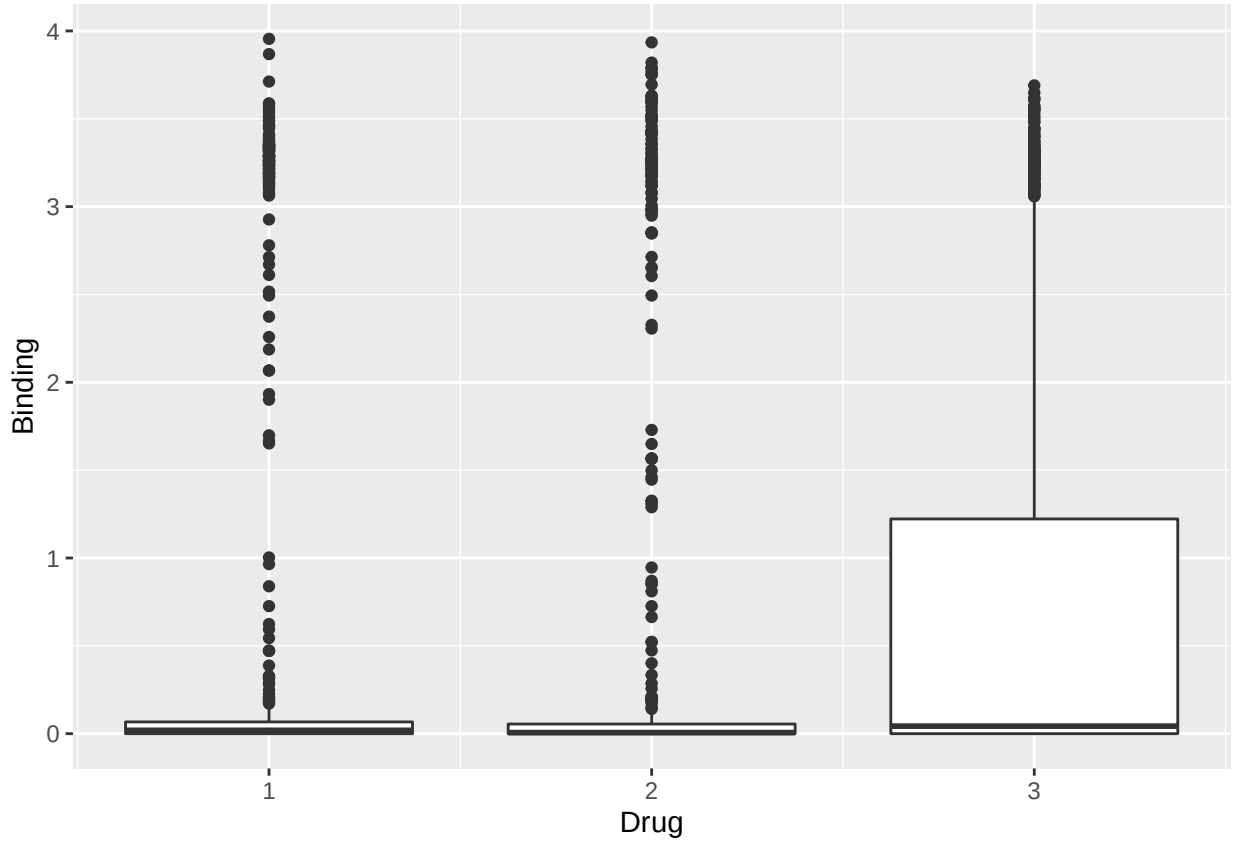
```
##
##     group 1 group 2 group 3 group 4 group 5 group 6 group 7
##   1     582     170      96       0       0       0       0
##   2       0       0       0     444     374     131       0
##   3       0       0       0       0       0       0     668
```

Now we look at the table of each marcaque and the corresponding treatment group. As the table shows, each marcaque only received one kidn of treatment.

```
table(Data2$MonkeyID, Data2$Treatment)
```

```
##
##        group 1 group 2 group 3 group 4 group 5 group 6 group 7
##   6104       0       0      35       0       0       0       0
##   6105       0       0       0     228       0       0       0
##   6107       0       0       0       0       0       0     239
##   6117     243       0       0       0       0       0       0
##   6118       0       7       0       0       0       0       0
##   6119       0       0      55       0       0       0       0
##   6125       0       0       0     216       0       0       0
##   6132       0       0       0       0     251       0       0
##   6160     183       0       0       0       0       0       0
##   6193       0     117       0       0       0       0       0
##   6199       0       0       0       0       0      48       0
##   6200       0       0       0       0       0       0     191
##   6201       0       0       0       0      73       0       0
##   6202       0       0       0       0       0      78       0
##   6203       0       0       0       0       0       0     238
##   6204     156       0       0       0       0       0       0
##   6205       0       0       0       0       0       5       0
##   6209       0      46       0       0       0       0       0
##   6210       0       0       0       0      50       0       0
##   6214       0       0       6       0       0       0       0
```

```
ggplot(Data2, aes(x = Drug, y = Binding)) + geom_boxplot(aes(group = Drug))
```




```
table(Data$Drug, Data$Reactivity)
```

```
##
##       0   1
##   1 680 168
##   2 807 142
##   3 464 204
```

```
table(Data$Time_Point, Data$Reactivity)
```

```
##
##       0   1
##   0 249  24
##   1 971  33
##   2 475 348
##   3 256 109
```

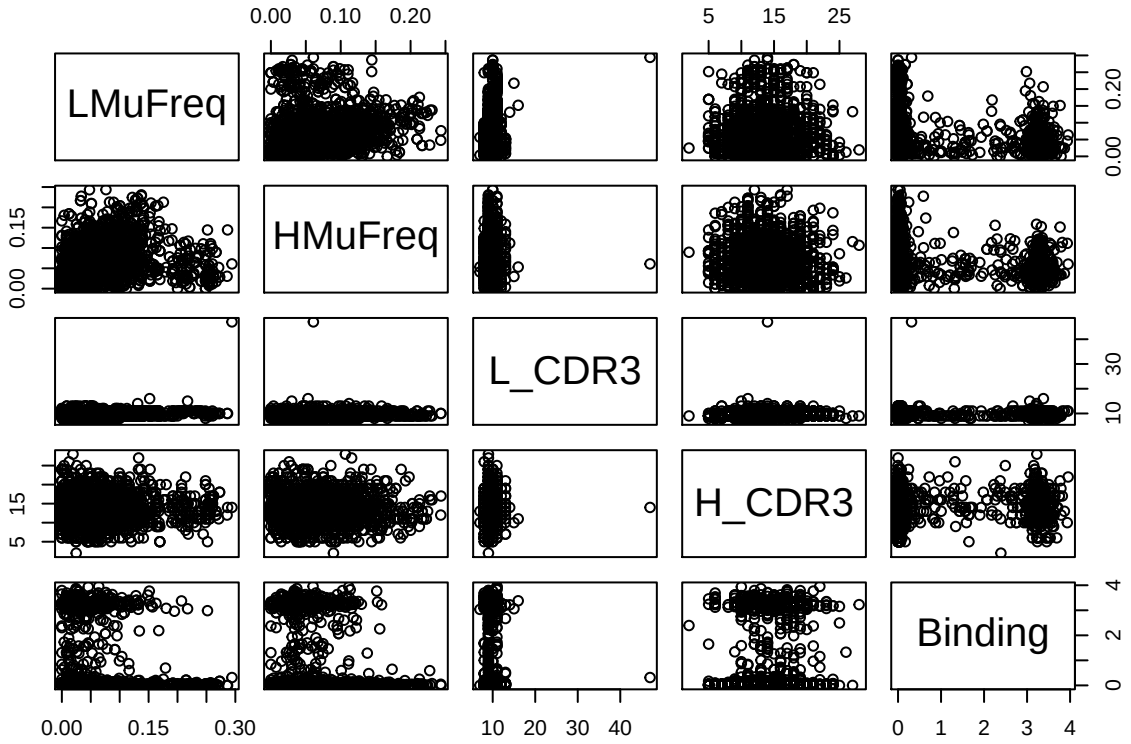
```
Data2 %>% group_by(Drug) %>% summarize(avgLMuFreq = mean(LMuFreq), avgHMuFreq = mean(HMuFreq), avgBindi
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 4
##    Drug avgLMuFreq avgHMuFreq avgBinding
##   <dbl>      <dbl>      <dbl>      <dbl>
## 1     1     0.0616    NA           0.450
## 2     2     0.0616     0.0730      0.334
```

```
## 3      3      0.0594      0.0559       0.807
```

```
Data2 %>% select(LMuFreq, HMuFreq, L_CDR3, H_CDR3, Binding) %>% pairs()
```



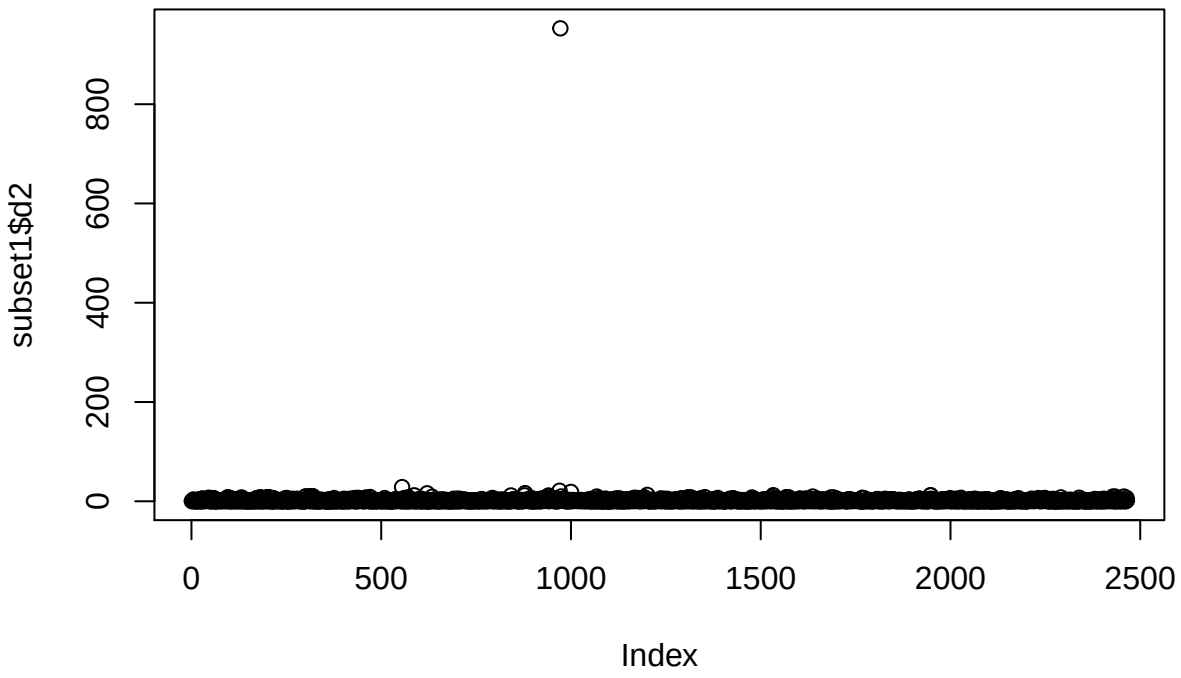
Outlier detection

Notice may have outlier in LCDR3 variable.

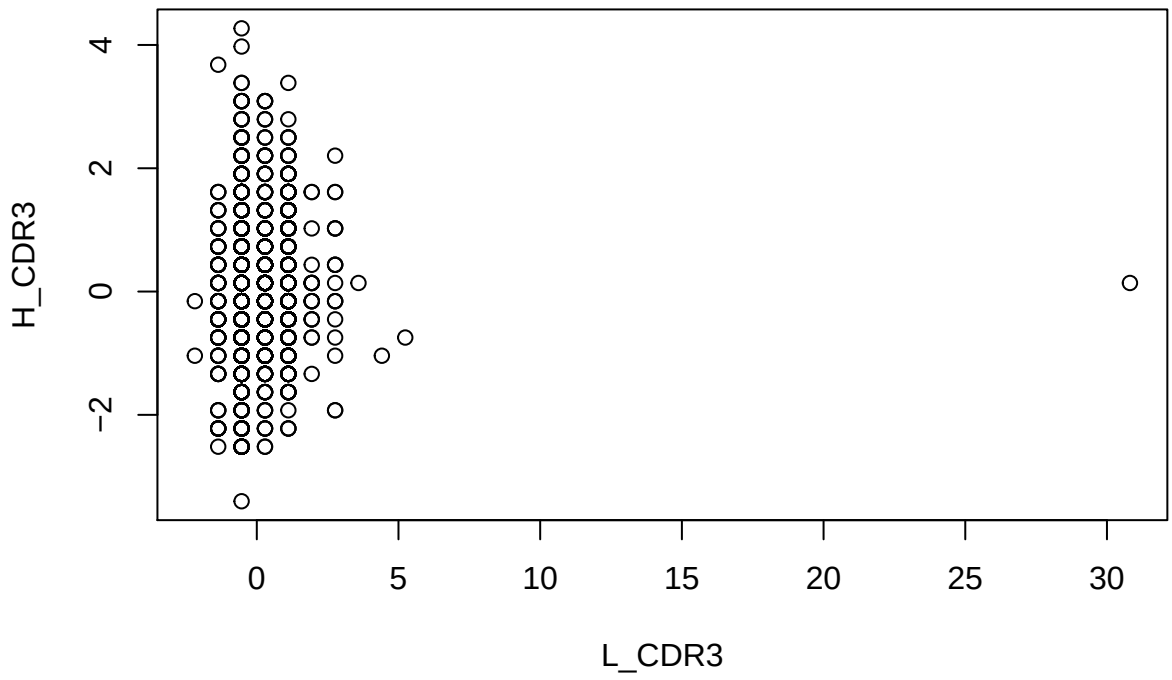
```
summary(Data2$L_CDR3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00    9.00    9.00    9.65   10.00   47.00
```

```
subset1 <- Data2 %>% select(L_CDR3, H_CDR3)
subset1$d2 <- mahalanobis(subset1, colMeans(subset1), cov(subset1))
subset1$Z <- scale(subset1)
plot(subset1$d2)
```




```
plot(subset1$Z)
```




```
subset2 <- subset1 %>% arrange(desc(d2), desc(Z))
subset2[1,]
```

```
## # A tibble: 1 x 4
##   L_CDR3 H_CDR3    d2 Z[,"L_CDR3"] [,"H_CDR3"] [,"d2"]
##    <dbl>  <dbl> <dbl>        <dbl>       <dbl>   <dbl>
## 1     47     14  953.         30.8       0.139    49.4
```

```
which(subset1$L_CDR3 == 47)
```

```
## [1] 972
```

Row 972 from `Data2` is in fact an outlier, as shown in the summary and plots above. The value for `L_CDR3` is quite unlikely. Since we can't go back to the original data, we remove the data point and will use the new dataset `Data3`.

```
Data2[972,]
```

```
## # A tibble: 1 x 24
##   MonkeyID  Drug Treatment Time_Point Isotype HV_Extract HD_Extract HJ_Extract
##      <dbl> <dbl> <chr>          <dbl> <chr>   <chr>      <chr>      <chr>
## 1     6107     3 group 7            2 G       3          2          4
## # ... with 16 more variables: H_VBase <dbl>, H_Substitutions <dbl>,
## #   H_Insertions <dbl>, H_Deletions <dbl>, HMuFreq <dbl>, H_CDR3 <dbl>,
## #   LV_Extract <chr>, LJ_Extract <chr>, L_VBase <dbl>, L_Substitutions <dbl>,
## #   L_Insertions <dbl>, L_Deletions <dbl>, LMuFreq <dbl>, L_CDR3 <dbl>,
## #   Binding <dbl>, Reactivity <dbl>
```

```r
Data3 <- Data2[-972,]
```

Multivariate Data Analysis

```r
ID <- as.factor(Data3$MonkeyID)
trt <- as.factor(Data3$Treatment)
drug <- as.factor(Data3$Drug)
tp <- as.factor(Data3$Time_Point)
it <- as.factor(Data3$Isotype)
# four-way manova


fit.manova <- manova(cbind(Data3$L_CDR3, Data3$LMuFreq, Data3$H_CDR3, Data3$HMuFreq, Data3$Binding) ~ d
summary(fit.manova)
```

```
##             Df  Pillai approx F num Df den Df    Pr(>F)
## drug         2 0.06176   15.626     10   4904 < 2.2e-16 ***
## tp           3 0.19800   34.667     15   7359 < 2.2e-16 ***
## Residuals 2455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fit.gls <- lm(cbind(Data3$L_CDR3, Data3$LMuFreq, Data3$H_CDR3, Data3$HMuFreq, Data3$Binding) ~ drug + t
summary(fit.gls)
```

```
## Response Data3$L_CDR3 :
##
## Call:
## lm(formula = `Data3$L_CDR3` ~ drug + tp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5951 -0.6553 -0.5665  0.4049  6.2942
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.65526    0.06178 156.281   <2e-16 ***
## drug2       -0.04384    0.04543  -0.965    0.335
## drug3       -0.11063    0.05064  -2.184    0.029 *
## tp1          0.01725    0.06568   0.263    0.793
## tp2          0.05051    0.06748   0.749    0.454
## tp3          0.02183    0.07682   0.284    0.776
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9508 on 2455 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.002056,   Adjusted R-squared:  2.318e-05
## F-statistic: 1.011 on 5 and 2455 DF,  p-value: 0.4092
##
##
## Response Data3$LMuFreq :
##
```

```
## Call:
## lm(formula = `Data3$LMuFreq` ~ drug + tp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.07233 -0.03662 -0.01504  0.01987  0.22320
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0719614  0.0035016  20.551  < 2e-16 ***
## drug2       -0.0001371  0.0025748  -0.053  0.95752
## drug3        0.0003640  0.0028703   0.127  0.89910
## tp1         -0.0084782  0.0037224  -2.278  0.02284 *
## tp2         -0.0147355  0.0038247  -3.853  0.00012 ***
## tp3         -0.0189684  0.0043541  -4.356 1.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05389 on 2455 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.01006,    Adjusted R-squared:  0.008045
## F-statistic:  4.99 on 5 and 2455 DF,  p-value: 0.0001496
##
##
## Response Data3$H_CDR3 :
##
## Call:
## lm(formula = `Data3$H_CDR3` ~ drug + tp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2921  -2.1012  -0.1012   1.8988  14.7079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.95536    0.21869  63.813  < 2e-16 ***
## drug2       -0.40087    0.16080  -2.493  0.01273 *
## drug3       -0.57162    0.17926  -3.189  0.00145 **
## tp1         -0.45325    0.23248  -1.950  0.05134 .
## tp2         -0.09164    0.23887  -0.384  0.70128
## tp3          0.69302    0.27193   2.549  0.01088 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.366 on 2455 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.01637,    Adjusted R-squared:  0.01436
## F-statistic:  8.17 on 5 and 2455 DF,  p-value: 1.162e-07
##
##
## Response Data3$HMuFreq :
##
## Call:
## lm(formula = `Data3$HMuFreq` ~ drug + tp)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.085072 -0.029604 -0.006174  0.024701  0.174404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.079295   0.002713  29.230  < 2e-16 ***
## drug2        0.005776   0.001995   2.896 0.003816 **
## drug3       -0.008182   0.002224  -3.679 0.000239 ***
## tp1         -0.010457   0.002884  -3.626 0.000294 ***
## tp2         -0.016569   0.002963  -5.592 2.50e-08 ***
## tp3         -0.021684   0.003373  -6.428 1.55e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04175 on 2455 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.04463,    Adjusted R-squared:  0.04268
## F-statistic: 22.94 on 5 and 2455 DF,  p-value: < 2.2e-16
##
##
## Response Data3$Binding :
##
## Call:
## lm(formula = `Data3$Binding` ~ drug + tp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1488 -0.8238 -0.0266  0.0164  3.2987
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.028715   0.065234   0.440    0.660
## drug2       -0.028075   0.047967  -0.585    0.558
## drug3        0.074364   0.053473   1.391    0.164
## tp1          0.009551   0.069347   0.138    0.890
## tp2          1.045709   0.071253  14.676   <2e-16 ***
## tp3          0.751695   0.081115   9.267   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.004 on 2455 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.1975, Adjusted R-squared:  0.1958
## F-statistic: 120.8 on 5 and 2455 DF,  p-value: < 2.2e-16
```

```
fit.logit <- lm(Data3$Reactivity ~ drug*tp)
summary(fit.logit)
```

```
##
## Call:
## lm(formula = Data3$Reactivity ~ drug * tp)
##
## Residuals:
```

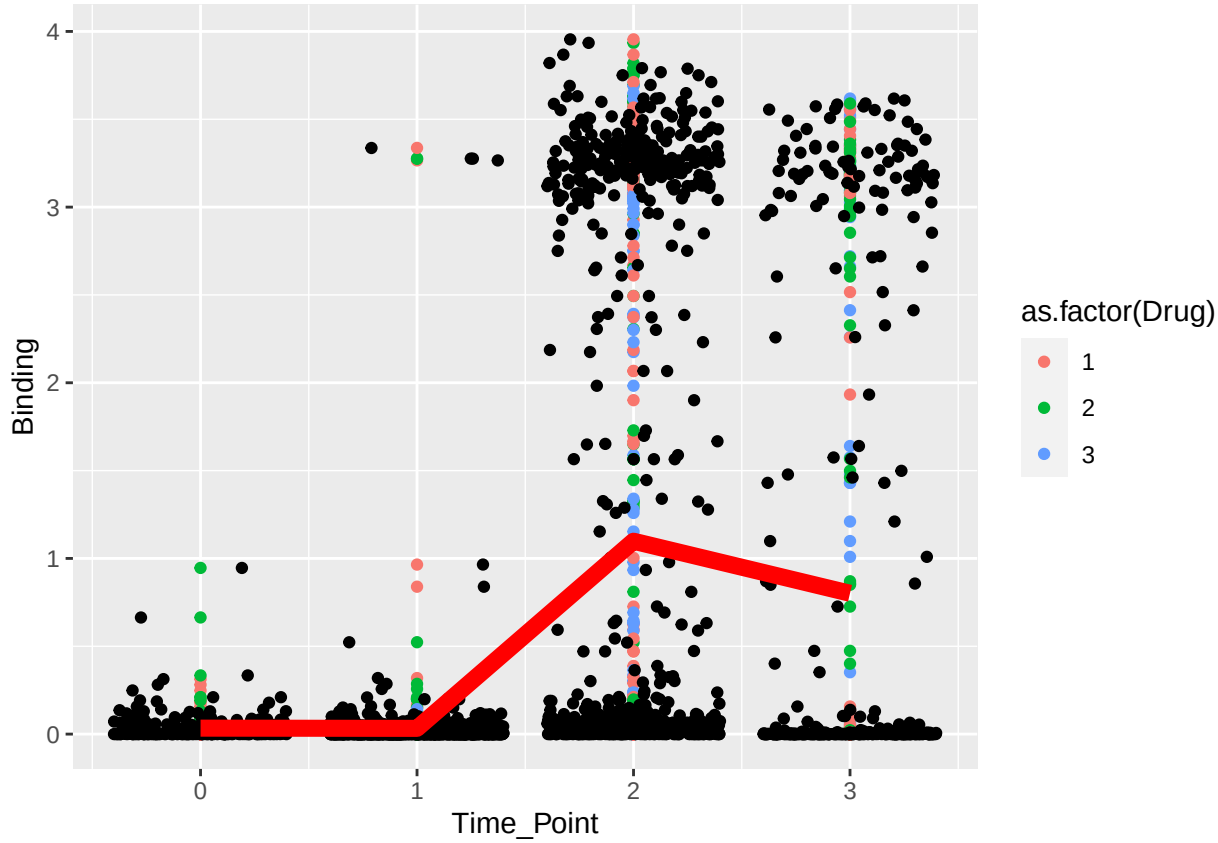
```
##      Min       1Q   Median       3Q      Max
## -0.45954 -0.30282 -0.04624 -0.00800  0.99200
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.108527   0.032132   3.378 0.000743 ***
## drug2        0.002584   0.050123   0.052 0.958890
## drug3       -0.108527   0.059152  -1.835 0.066668 .
## tp1         -0.062284   0.037649  -1.654 0.098182 .
## tp2          0.333704   0.039536   8.440  < 2e-16 ***
## tp3          0.112784   0.046089   2.447 0.014471 *
## drug2:tp1   -0.018808   0.056100  -0.335 0.737458
## drug3:tp1    0.070284   0.070352   0.999 0.317874
## drug2:tp2   -0.102593   0.060290  -1.702 0.088949 .
## drug3:tp2    0.125834   0.066442   1.894 0.058357 .
## drug2:tp3    0.162243   0.070163   2.312 0.020839 *
## drug3:tp3    0.190033   0.074355   2.556 0.010655 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.365 on 2452 degrees of freedom
## Multiple R-squared:  0.196,  Adjusted R-squared:  0.1924
## F-statistic: 54.34 on 11 and 2452 DF,  p-value: < 2.2e-16
```

Longitudinal Data Analysis

First we don't consider treatments but only plot the mean trend over time.

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

```
# simply connects the mean of each time point
ggplot(Data3, aes(x = Time_Point, y = Binding)) + geom_point(aes(color = as.factor(Drug))) + geom_jitte
```



Here we use Binding as the only response. Prdictors: `Drug`.
Random effect for both intercept and slope.

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} + e_{ij}$$

```
lda <- lme(fixed = Binding ~ Time_Point + Drug,
           random = ~ Time_Point | MonkeyID, data = Data3, method = "REML")
summary(lda)
```

```
## Linear mixed-effects model fit by REML
##  Data: Data3
##        AIC      BIC    logLik
##   6738.894 6779.552 -3362.447
##
## Random effects:
##  Formula: ~Time_Point | MonkeyID
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev    Corr
## (Intercept) 0.6043473 (Intr)
## Time_Point  0.5938004 -0.97
## Residual    0.9330025
##
## Fixed effects: Binding ~ Time_Point + Drug
##                  Value  Std.Error   DF   t-value p-value
## (Intercept) -0.5671081 0.21736533 2443 -2.609009  0.0091
## Time_Point   0.6404275 0.18256794 2443  3.507886  0.0005
```

```
## Drug         0.0414165 0.05945489   18  0.696604  0.4949
##  Correlation:
##            (Intr) Tm_Pnt
## Time_Point -0.847
## Drug       -0.499  0.007
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.52231270 -0.51818636 -0.09479562  0.01991047  3.61431329
##
## Number of Observations: 2464
## Number of Groups: 20
```