Modeling Treatment Effects Using Multilevel Analysis

Shu-Yu, Lin

March 31,2025

```
invisible(
  lapply(c("lme4", "lmerTest", "ggplot2", "sjPlot","effectsize", "performance",
           "gridExtra", "patchwork","dplyr", "ggeffects", "broom.mixed",
           "ggpubr"),
         require, character.only = TRUE))
```

```
data <- read.csv("https://danieleturchetti.github.io/MATH43515/hospital.csv",header=TRUE)
head(data)
```

```
##   Patient_ID Hospital_ID      Age Severity    Treatment Time Improvement
## 1          1           7 54.54917 7.649358 Experimental    1     5.49614
## 2          1           7 54.54917 7.649358 Experimental    2    21.20246
## 3          1           7 54.54917 7.649358 Experimental    3    11.77205
## 4          1           7 54.54917 7.649358 Experimental    4    22.09269
## 5          2           4 58.19140 8.381224 Experimental    1    12.95645
## 6          2           4 58.19140 8.381224 Experimental    2    30.14875
```

```
dim(data)
```

```
## [1] 1000    7
```

Introduction (15 marks)

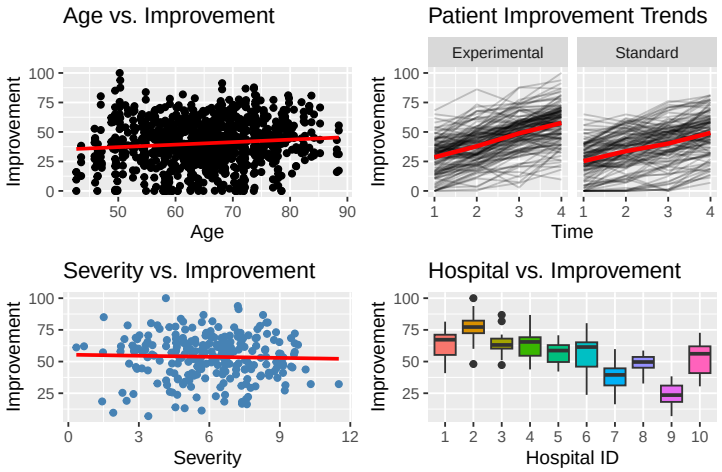
The dataset used in this analysis pertains to patient improvement across different treatment types, hospitals, and over time. Each patient has multiple observations, making this suitable for hierarchical or multilevel modelling. The primary variable of interest is "Improvement," which quantifies patient recovery or improvement after treatment. We aim to explore factors influencing improvement, including patient age, severity of condition, type of treatment administered, and variations across different hospitals.

Exploratory visualizations support the multilevel structure and provide insights into variable relationships. Scatter plots of age versus improvement suggest a weak relationship, implying that age alone may not be a strong predictor. Likewise, the relationship between severity and improvement is not visually prominent.Line plots tracking improvement over time show that most patients experience some level of recovery, with the experimental group demonstrating more significant gains overall. This trend is further supported by plots incorporating average trajectories. Lastly, boxplots of improvement across hospitals highlight differences in patient outcomes that point to the importance of accounting for hospital-level effects. These findings motivate multilevel modelling to examine treatment effects, patient characteristics, and hospital variability more rigorously.

Although exploratory data analysis (EDA) suggested that variables such as age and severity showed a limited direct association with improvement, we included them in the multilevel modelling phase to allow

for the possibility that their effects may emerge in the presence of interactions or after accounting for higher-level structure. Specifically, these variables may have context-dependent influences, or their effects may be moderated by treatment type, time, or group-level variation. Incorporating them into the hierarchical model allows for a more rigorous assessment of their potential role in predicting patient improvement.

```
# 1. Scatterplot: Age vs. Improvement
p1 <- ggplot(data, aes(x=Age, y=Improvement)) +
  geom_point(color="black") +
  geom_smooth(method="lm", se=FALSE, color="red") +
  labs(title="Age vs. Improvement")
# 2. Line plot with mean line
p2 <- ggplot(data, aes(x = Time, y = Improvement, group = Patient_ID)) +
  geom_line(alpha = 0.2) +
  stat_summary(aes(group = 1), fun = mean, geom = "line", color = "red",
               linewidth = 1.2) +
  facet_wrap(~Treatment) +
  labs(title = "Patient Improvement Trends")
# 3. Scatterplot: Severity vs. Improvement
last_improvement <- data %>%
  group_by(Patient_ID) %>%
  filter(Time == max(Time)) %>%
  ungroup()
p3 <- ggplot(last_improvement, aes(x = Severity, y = Improvement)) +
  geom_point(color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Severity vs. Improvement")
# 4. Hospital factor analysis, compare the average improvement across different hospitals
p4 <- ggplot(last_improvement, aes(x = as.factor(Hospital_ID), y = Improvement,
                                   fill = as.factor(Hospital_ID))) +geom_boxplot() +
  labs(title = "Hospital vs. Improvement", x = "Hospital ID", y = "Improvement")+
  theme(legend.position = "none")
print(group1 <- ( p1 | p2 ) / ( p3 | p4 ))
```



Methods (25 marks)

Multilevel models, also known as hierarchical linear models or mixed-effects models, are statistical approaches designed to handle data with nested or clustered structures. In the context of this dataset, patients are

nested within hospitals and are measured repeatedly over time. This structure violates the assumption of independence required in traditional regression models because measurements from the same patient (or hospital) are likely to be correlated. Multilevel models address this issue by allowing for random effects at different grouping levels, thus partitioning the variability attributable to patients, hospitals, and residual error.

Given the structure of our data, we employ a three-level longitudinal multilevel modelling framework. The first level consists of repeated measures across time for each patient. The second level captures patient variation and the third level accounts for differences between hospitals. This approach enables us to model within-patient improvement over time while accounting for both patient- and hospital-level clustering.

To better understand how much variation in improvement is attributable to each level, we decompose the total variance using intraclass correlation coefficients (ICC). The ICC quantifies the proportion of variance explained by higher-level groupings (e.g., hospital and patient). We also compute the variance partition coefficient (VPC) to evaluate how much total variation is accounted for at each level. These statistics guide our understanding of the need for multilevel modelling and help justify the inclusion of random intercepts (and potentially slopes) for hospital and patient effects.

Model building is performed incrementally. We begin with a simple model that examines the overall effect of treatment on improvement without accounting for grouping. Subsequently, we add random effects for hospitals and patients to reflect the nested structure. Finally, we incorporate time and potential interaction effects (e.g., Treatment * Time) to capture longitudinal patterns. Model comparison is conducted at each step using likelihood ratio tests, AIC, and BIC values. Fixed effect estimates, random effect variances, and residual diagnostics are examined to ensure model validity and fit.

Analysis (35 marks)

This section presents the step-by-step model-building and analysis process to investigate the factors associated with patient improvement. Starting from a simple model examining the direct effect of treatment, we gradually incorporate additional levels of the data hierarchy—such as hospital clustering and repeated measures over time—and control variables like age and severity.Each model is carefully constructed to address specific research questions, including the effectiveness of different treatments, hospital-level heterogeneity, time-dependent effects, and the influence of other patient-level characteristics. At each stage, appropriate statistical tests, model comparison metrics (e.g., AIC, likelihood ratio tests), and diagnostic tools (e.g., residual plots) are applied to assess the quality and assumptions of the models.

Model 1

We first assessed the overall treatment effect using a simple linear regression without accounting for clustering. Results show that patients receiving the experimental treatment had significantly higher improvement scores than those receiving the standard treatment ($\beta = -6.12$, $p < 0.001$). The estimated effect size was moderate (Cohen's $d = 0.32$, 95% CI [0.20, 0.45]), indicating a practically meaningful advantage for the experimental group. Although this model does not account for the hierarchical structure of the data, it provides an initial confirmation of treatment effectiveness.

```
# Build the model: Improvement ~ Treatment
model1 <- lm(Improvement ~ Treatment, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = Improvement ~ Treatment, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.223 -12.652   0.644  13.557  56.777
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        43.2229     0.8239  52.458  < 2e-16 ***
## TreatmentStandard  -6.1219     1.2096  -5.061 4.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.08 on 998 degrees of freedom
## Multiple R-squared:  0.02502,    Adjusted R-squared:  0.02405
## F-statistic: 25.61 on 1 and 998 DF,  p-value: 4.961e-07
```

```
#tidy(model1, effects = "fixed")
confint(model1)
```

```
##                       2.5 %    97.5 %
## (Intercept)       41.606006 44.839744
## TreatmentStandard -8.495537 -3.748245
```

```
# Calculate Cohen's d to measure effect size between treatment groups
cohens_d(Improvement ~ Treatment, data = data)
```

```
## Cohen's d |       95% CI
## ------------------------
## 0.32      | [0.20, 0.45]
##
## - Estimated using pooled SD.
```

Model 2

To account for hospital-level variation in baseline improvement, we fitted a mixed-effects model with Treatment as a fixed effect and Hospital_ID as a random intercept. This allows each hospital to have its own baseline level of improvement while assuming a constant treatment effect across hospitals. Results show that the Experimental group still demonstrated significantly higher improvement than the Standard group ($\beta = -5.62$, $p < 0.001$), confirming the treatment effect remains robust after adjusting for between-hospital differences. The standard deviation of the random intercept for hospitals was 14.42, compared to a residual standard deviation of 13.78, indicating substantial between-hospital variability. This is further supported by the Intraclass Correlation Coefficient (ICC) of 0.523, suggesting that over 52% of the variance in improvement can be attributed to hospital-level effects. To test whether the treatment effect varies across hospitals, we fitted a second model allowing random slopes of Treatment (model2.1). However, model comparison showed no significant improvement ($\chi^2(2) = 0.11$, $p = 0.946$), and visualizations showed only minor deviations across hospitals. Thus, the simpler random intercept model was sufficient to capture institutional differences without overfitting. These findings highlight the importance of accounting for hierarchical clustering in modeling treatment outcomes and justify the inclusion of hospital-level random effects in further multilevel modeling.

```
# Mixed effects model: Treatment as fixed effect, Hospital as random effect
model2 <- lmer(Improvement ~ Treatment + (1 | Hospital_ID), data = data)
model2.1 <- lmer(Improvement ~ Treatment + (Treatment | Hospital_ID),data=data)
anova(model2,model2.1)
```

```
## Data: data
## Models:
## model2: Improvement ~ Treatment + (1 | Hospital_ID)
## model2.1: Improvement ~ Treatment + (Treatment | Hospital_ID)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2      4 8136.4 8156.0 -4064.2   8128.4
## model2.1    6 8140.3 8169.7 -4064.1   8128.3 0.1113  2     0.9459
```

```
VarCorr(model2)
```

```
##  Groups      Name        Std.Dev.
##  Hospital_ID (Intercept) 14.418
##  Residual                13.775
```

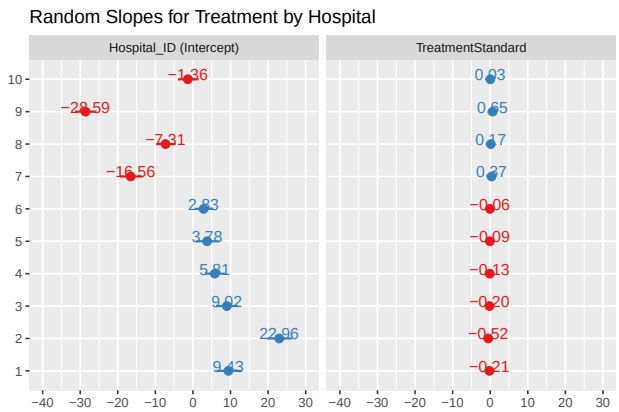
```
tidy(model2, effects = "fixed")
```

```
## # A tibble: 2 x 7
##   effect term              estimate std.error statistic     df  p.value
##   <chr>  <chr>                <dbl>     <dbl>     <dbl>  <dbl>    <dbl>
## 1 fixed  (Intercept)          43.9      4.60       9.55   9.14 4.71e- 6
## 2 fixed  TreatmentStandard    -5.62     0.886     -6.34 990.   3.42e-10
```

```
# Calculate Intraclass Correlation Coefficient (ICC)
icc(model2)
```

```
## # Intraclass Correlation Coefficient
##
##     Adjusted ICC: 0.523
##   Unadjusted ICC: 0.513
```

```
# Visualize the random slopes of Treatment by hospital
plot_model(model2.1, type = "re", sort.est = TRUE, show.values = TRUE,
           title = "Random Slopes for Treatment by Hospital")
```



Model 3

To examine how treatment effects vary over time while accounting for clustering at both hospital and patient levels, we fit a mixed-effects model with random intercepts for both Hospital_ID and Patient_ID, and

included a Treatment × Time interaction term. Incorporating a patient-level random intercept is crucial when evaluating the effect of Time, as repeated measurements are taken on the same individuals. Without modeling within-patient correlation, we risk violating the independence assumption of residuals, which can lead to biased standard errors and incorrect inferences. By allowing each patient to have their own baseline (random intercept), the model captures individual variability in initial improvement levels. This is particularly important when assessing longitudinal change (Time), as patients may start at different points and improve at different rates. The inclusion of the patient-level intercept thus ensures more accurate estimation of the Time effect and its interaction with Treatment, while properly accounting for the nested structure of the data. The interaction was statistically significant ($p < 0.001$), suggesting that the improvement trajectory differed by treatment group. Specifically, the Experimental group improved more rapidly than the Standard group over time. Model comparison with the additive model (without the interaction term) showed a significantly better fit ($\Delta AIC = 22.8$; $p < 0.001$), supporting the inclusion of the interaction. Among the fixed effects, Time was a strong positive predictor (estimate = 9.82, $p < 0.001$), while the treatment effect alone (TreatmentStandard) was not significant ($p = 0.635$), further emphasizing the importance of the interaction. Variance partitioning revealed that approximately 70% of the total variance was attributable to hospital-level (13.8%) and patient-level (55.9%) differences. The high Intraclass Correlation Coefficients (ICC_hospital = 0.14; ICC_patient = 0.56) justify the use of a multilevel framework. Overall, this model highlights that treatment effects evolve over time, and that substantial between-patient and between-hospital variability must be accounted for in evaluating treatment outcomes.

```
# Random intercepts for both hospitals and patients (each patient has a different baseline)
model3 <- lmer(
  Improvement ~ Treatment + Time + (1 | Hospital_ID) + (1 | Patient_ID),
  data = data)
model3.1 <- lmer(
  Improvement ~ Treatment * Time + (1 | Hospital_ID) + (1 | Patient_ID),
  data = data)
anova(model3,model3.1)
```

```
## Data: data
## Models:
## model3: Improvement ~ Treatment + Time + (1 | Hospital_ID) + (1 | Patient_ID)
## model3.1: Improvement ~ Treatment * Time + (1 | Hospital_ID) + (1 | Patient_ID)
##          npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3      6 7176.4 7205.8 -3582.2   7164.4
## model3.1    7 7153.6 7188.0 -3569.8   7139.6 24.724  1  6.615e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
VarCorr(model3.1)
```

```
##  Groups      Name        Std.Dev.
##  Patient_ID  (Intercept)  6.3980
##  Hospital_ID (Intercept) 14.4086
##  Residual                 7.0052
```

```
tidy(model3.1, effects = "fixed")
```

```
## # A tibble: 4 x 7
##   effect term                  estimate std.error statistic     df   p.value
##   <chr>  <chr>                    <dbl>     <dbl>     <dbl>  <dbl>     <dbl>
## 1 fixed  (Intercept)              19.4       4.65      4.16    9.56 2.14e-  3
```

```
## 2 fixed  TreatmentStandard        -0.648     1.37     -0.474 767.   6.35e-  1
## 3 fixed  Time                      9.82      0.271    36.3   748.   5.38e-167
## 4 fixed  TreatmentStandard:Time   -1.99      0.397    -5.01  748.   6.90e-  7
```

```
REsummary <- as.data.frame(VarCorr(model3.1))
sig <- REsummary$vcov[3] #Residual variance
sigv <- REsummary$vcov[2] # variance for Hospital
sigu <- REsummary$vcov[1] # variance for Patient
totalvar <- sum(REsummary$vcov) #total variance
vpc.hospital <- sigv/totalvar
vpc.patient <- sigu/totalvar
icc.hospital <- sigv/totalvar
icc.patient <- (sigu+sigv)/totalvar
print(c(vpc.hospital,vpc.patient,icc.hospital,icc.patient))
```

```
## [1] 0.6975752 0.1375400 0.6975752 0.8351153
```

Model 4

A full model was constructed to explore whether additional covariates help explain the treatment effects, including Age and Severity, along with the Treatment × Time interaction. Using a likelihood ratio test, this model was compared against the previous three-level random intercept model. The result (p = 0.3934) indicated that including Age and Severity did not significantly improve model fit. This finding aligns with insights from the exploratory analysis, where Age and Severity showed limited direct association with Improvement. Nonetheless, including them in the model allowed us to verify that their effects are not substantial even after accounting for interactions and nested structure, reinforcing the robustness of the simpler model.

```
model_full <- lmer(Improvement ~ Treatment * Time + Age + Severity +
                   (1 | Hospital_ID) + (1 | Patient_ID), data = data)
tidy(model_full, effects = "fixed")
```

```
## # A tibble: 6 x 7
##   effect term                  estimate std.error statistic    df   p.value
##   <chr>  <chr>                    <dbl>     <dbl>     <dbl> <dbl>     <dbl>
## 1 fixed  (Intercept)             17.2       5.90       2.92   24.7 7.29e-  3
## 2 fixed  TreatmentStandard       -0.685     1.37      -0.501 762.  6.16e-  1
## 3 fixed  Time                     9.82      0.271     36.3   748.  5.38e-167
## 4 fixed  Age                      0.0518    0.0512     1.01  237.  3.12e-  1
## 5 fixed  Severity                -0.210     0.235     -0.893 237.  3.73e-  1
## 6 fixed  TreatmentStandard:Time  -1.99      0.397     -5.01  748.  6.90e-  7
```

Final Fitted Model

To comprehensively evaluate treatment effects over time while accounting for the hierarchical structure of the data, we selected the model3.1 as the final fitted model, which can be written as:

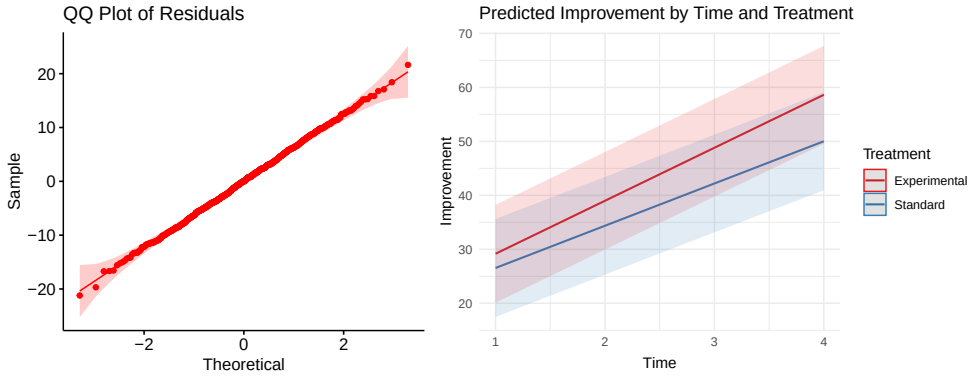
$$Yijk = \beta 0 + \beta 1 \cdot Treatmentij + \beta 2 \cdot Timek + \beta 3 \cdot (Treatmentij \times Timek) + uj + vi + \epsilon ijkYijk$$

This model includes random intercepts for both hospitals and patients, capturing unobserved heterogeneity at both levels. Compared to simpler models, this specification yielded significantly better model fit (ΔDeviance = 24.7, $p < 0.001$), indicating that accounting for individual patient baselines significantly improves

explanatory power. The model also explained a large proportion of variance at the hospital (VPC = 0.70) and patient level (VPC = 0.14), supporting the need for multilevel modeling.

Model diagnostics reveal a strong linear relationship between actual and predicted values, while the Q–Q plot indicates that residuals are approximately normally distributed, supporting key model assumptions. The significant interaction between Treatment and Time ($p < 0.001$) suggests that treatment effects vary over time. Predicted trajectories show that patients in the experimental group improve at a faster rate than those in the standard group, with the gap widening across time. These trajectories include 95% confidence intervals, depicted as shaded bands, which illustrate the uncertainty around the estimated effects. The narrow bands reinforce the robustness of the treatment effect over time. Overall, the diagnostics confirm model adequacy and support the interpretation of longitudinal improvement trends.

```
preds <- ggpredict(model3.1, terms = c("Time", "Treatment"))
p5 <- plot(preds) +
  labs(title = "Predicted Improvement by Time and Treatment") +
  theme_minimal()
p6 <- ggqqplot(resid(model3.1),
               title = "QQ Plot of Residuals",
               color = "red")
(p6 | p5)
```



Discussion of results (15 marks)

The final model revealed that the experimental treatment led to significantly faster improvement compared to the standard treatment, particularly as time progressed. The significant interaction term between Treatment and Time confirms that the treatment effect is time-dependent—patients receiving the experimental treatment improved more rapidly over the observed time points. This pattern was further visualized in the predicted trajectories, where the performance gap between treatment groups widened consistently across time. These results suggest switching to the experimental treatment for all patients may be worthwhile. Although age and severity were included in an extended model to test for potential confounding or moderating effects, neither variable significantly improved model fit. This finding aligns with initial exploratory results and supports the robustness of the simpler interaction model.

Some limitations must be acknowledged. The sample size was relatively modest (10 hospitals and 250 patients), which may affect generalizability. Additionally, the model assumes linear trajectories of improvement and does not incorporate other real-world considerations such as treatment cost or side effects. Nevertheless, the findings consistently demonstrate that the experimental treatment is more effective over time, providing strong evidence in favor of its broader adoption. Future research involving larger and more diverse populations, longer follow-up, or non-linear growth models would help confirm and extend these conclusions.