# PHP2517: Applied Multilevel Data Analysis

Final Exam

Antonella Basso

May 16, 2022

## Data

The "**birthwt.csv**" dataset includes data collected from a study of infant birthweight conducted by the CDC in Georgia between 1980 and 1992. The data corresponds to longitudinal birthweights (in grams) of babies born to 878 mothers who have 5 births recorded. [More information about the study can be found here: https://www.ncbi.nlm.nih.gov/pubmed/9054238]. The `birthwt.csv` dataset has information on the following variables:

- **mid**: Mother ID
- **cid**: Child ID
- **weight**: Weight of the child at birth (in grams)
- **order**: Birth order of the child (1 to 5 for each mother)
- **age**: Age of the mother at the time of birth (in years)
- **age_c**: Age of the mother at the time of birth (in years), centered at 21
- **age0**: Age of the mother at the time of first birth (in years), centered at 21
- **ageDiff**: Difference in age of the mother (in years) between first and current birth

**Notation**: For model specification in the questions of this exam we use the following,

- $i$: Mother index
- $j$: Child index
- $Y_{ij}$: Birth weight of the $j^{\text{th}}$ child born to the $i^{\text{th}}$ mother
- $A_{i1}$: Age of the $i^{\text{th}}$ mother at the time of birth of her first child ($j = 1$), centered at 21 years
- $A_{iD}$: Difference in age of the $i^{\text{th}}$ mother between the births of her first and $j^{\text{th}}$ child ($A_{ij} - A_{i1}$)

## Question 1: EDA

Conduct an Exploratory Data Analysis (EDA). Provide useful summary statistics, and optionally plots, to describe the information in the available dataset.

### Solution

Overview of Data:

- 4,390 total observations or unique children corresponding to 878 unique mothers
- 5 observations/children per mother[1]
- Birth weights range between 312-5,528 grams, with a mean of 3,156 grams
- Mother ages (non-centered) range between 12-42, with a mean of 21 years
- Age differences between a mother's first and fifth child range between 3-15, with a mean of 7 years

[1] Such that there are 878 observations in each child index group.

**Descriptive Statistics***:

Table 1: Descriptive Statistics for First 5 Mothers

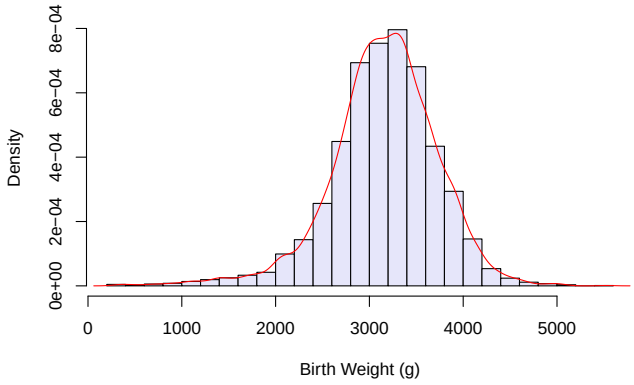
| | Primary Outcome | | | Other Predictors | |
|---|---|---|---|---|---|
| Mother Index | Children | Birth Weight Mean | Birth Weight Var | Mother's Age Mean | Mother's Age Var |
| 1 | 5 | 3634 | 192718 | 23 | 16 |
| 2 | 5 | 3340 | 302569 | 19 | 15 |
| 3 | 5 | 2506 | 185054 | 23 | 16 |
| 4 | 5 | 3018 | 284620 | 29 | 12 |
| 5 | 5 | 3062 | 72307 | 19 | 10 |

Table 2: Descriptive Statistics by Child Order

| | Primary Outcome | | | Other Predictors | | | |
|---|---|---|---|---|---|---|---|
| Child Index | Children | Birth Weight Mean | Birth Weight Var | Mother's Age Mean | Mother's Age Var | Age Diff. Mean | Age Diff. Var |
| 1 | 878 | 3099 | 280983 | 18 | 12 | 0 | 0 |
| 2 | 878 | 3142 | 302393 | 20 | 13 | 2 | 1 |
| 3 | 878 | 3173 | 332199 | 22 | 14 | 4 | 2 |
| 4 | 878 | 3167 | 361472 | 24 | 16 | 6 | 3 |
| 5 | 878 | 3201 | 345599 | 25 | 18 | 7 | 4 |

**EDA Plots**:

**Figue 1: Birth Weight Density**




**Figue 2: Centered Mother's Age Density**

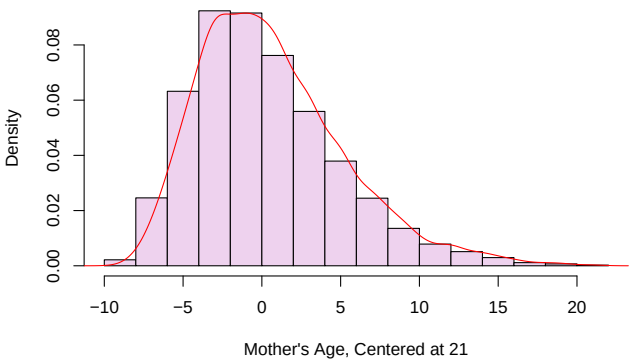



**Figure 3: Mother's Age Difference Density (Between Child 1−5)**

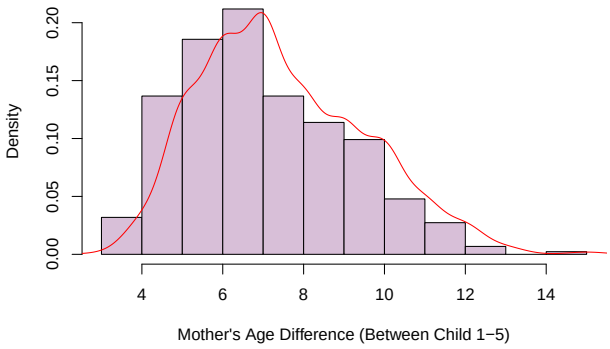


**Figure 4: Birth Weight Densities by Child Index**

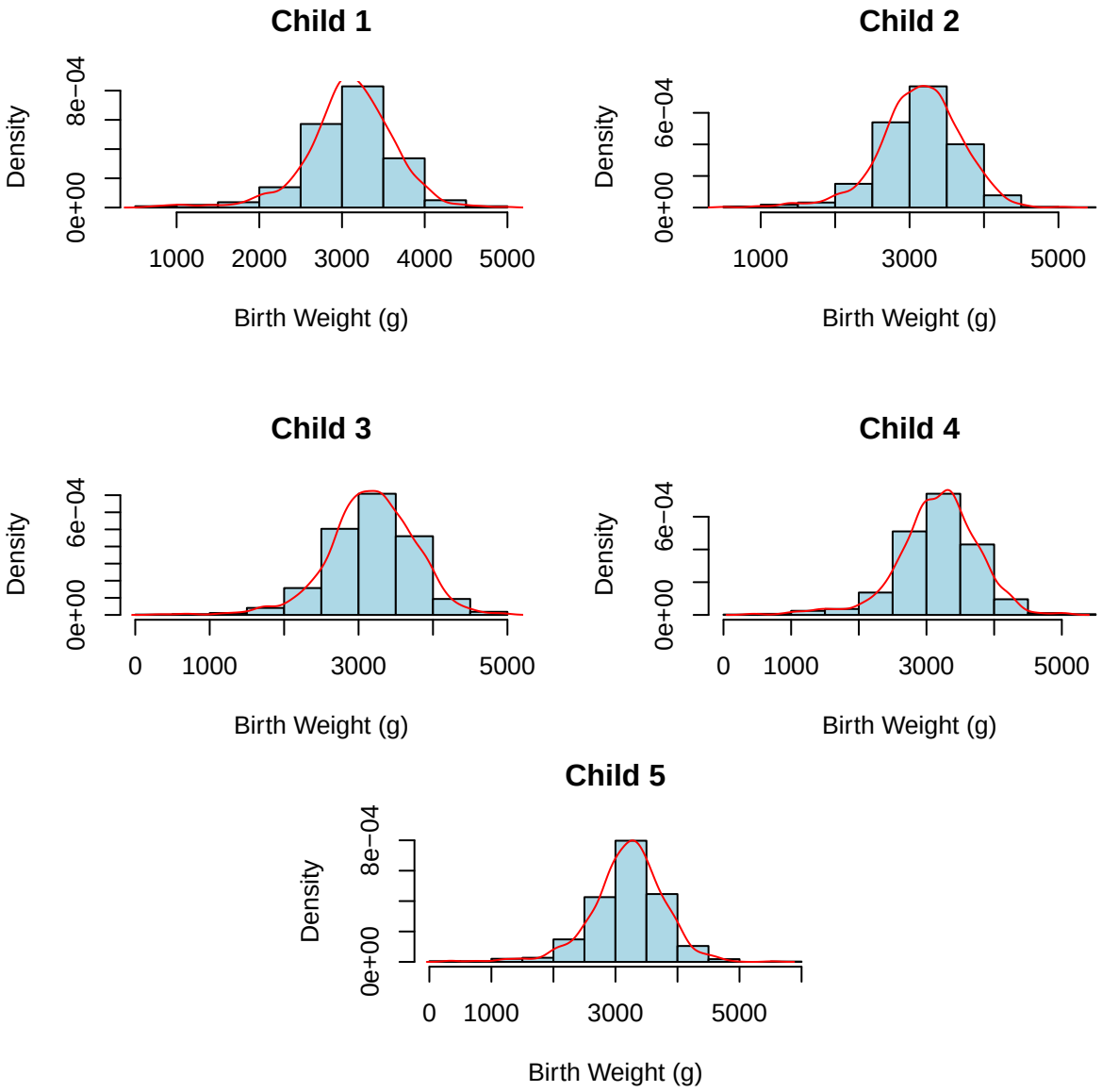


Figure 5: Spread of Birth Weight by Child Order

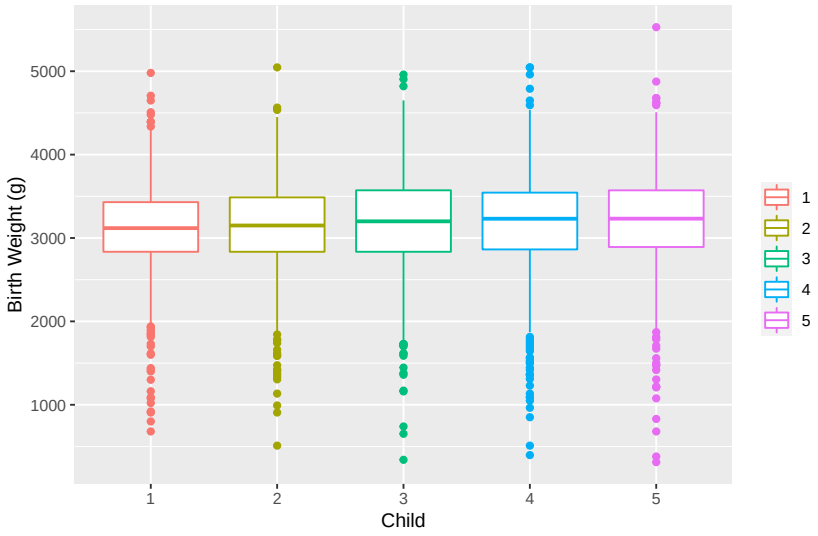



Figure 6: Birth Weight vs. Mother's Age by Child Order

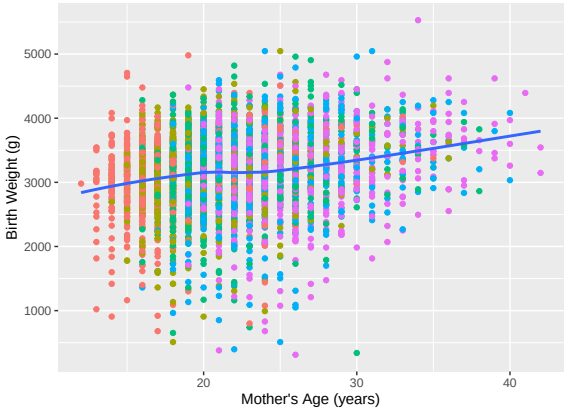



Figure 8: Mean Birth Weights by Mean Mother's Age

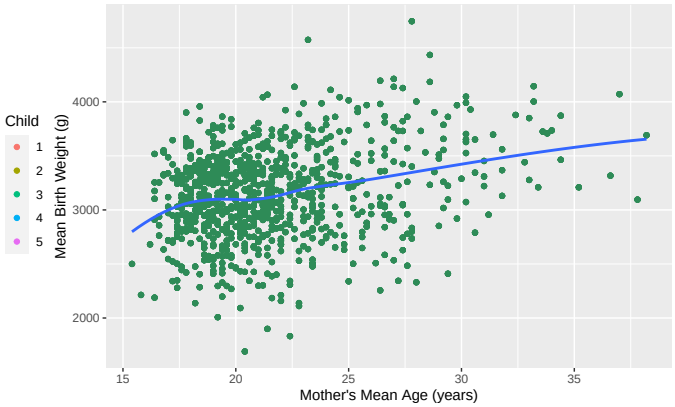
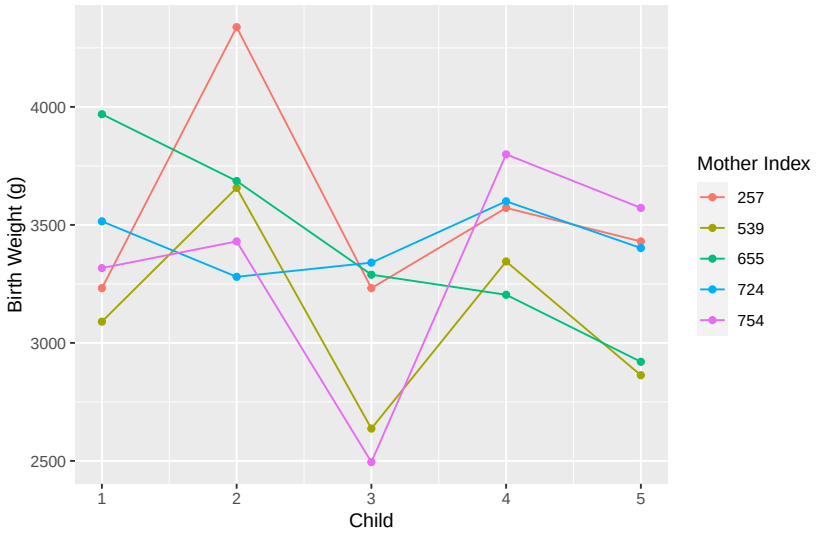



Figure 9: Birth Weight Trajectories for Randomly Selected Mothers



## Question 2:

Explain why multilevel statistical methods are appropriate for analyzing this data set. (1-2 sentences)

**Solution**

Multilevel statistical methods are appropriate for modeling this data in that observations display a clear hierarchical structure. Particularly, given that there are five children and hence, five birth weight observations corresponding to each mother, it is reasonable to model children as nested within mothers, as it is not uncommon to regard observations from biologically-tied individuals as being dependent. Moreover, since our outcome of interest can be described as a function of predictors, most of which are not constant across observations belonging to a single mother, it is also possible for individual observations to be correlated in such respects. Thus, multilevel analysis is appropriate to explore these intuitions, as it its primary objective is to model clustered data and provide insight into underlying patterns in the outcome at both a group and an individual level.

## Question 3:

Calculate estimates of mean birth weight using the following approaches:

a. Complete pooling
b. No-pooling
c. Partial-pooling

Compare the three means. What do you observe? Discuss the pros and cons of each approach.

**Solution**

a. **Complete Pooling**

Formula Estimate:

$$\bar{Y}_{\text{all}} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}}{\sum_{i=1}^{I} n_i} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} Y_{ij}}{N} = 3156.304 \approx 3156$$

Regression Estimate:

```
m_cp <- lm(weight ~ 1, data=birthwt)
m_cp$coef

(Intercept)
3156.304
```

b. **No-Pooling**

Formula Estimate:

$$\bar{\bar{Y}}_i = \frac{\sum_{i=1}^{I} \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}}{I} = \frac{\sum_{i=1}^{I} \bar{Y}_i}{I} = 3156.304 \approx 3156$$

Regression Estimate:

```
m_np <- lm(weight ~ -1 + mid, data=birthwt)
m_np2 <- lm(m_np$coef ~ 1)
m_np2$coef
```

```
(Intercept)
3156.304
```

c. **Partial-Pooling**

Formula Estimate:

$$\bar{Y}_{\text{pooled}} = \frac{\sum_{j=1}^{J}\left[\frac{\left(n_j/\hat{\sigma}_j^2\right)\bar{Y}_j+\left(1/\hat{\sigma}_\alpha^2\right)\bar{Y}_{\text{all}}}{n_j/\hat{\sigma}_j^2+1/\hat{\sigma}_\alpha^2}\right]}{I} = \frac{\sum_{j=1}^{J}\hat{a}_j}{I} = 3166.94 \approx 3157$$

Regression Estimate:

```
m_pp <- lmer(weight ~ 1 + (1|order), data=birthwt)
summary(m_pp)$coef

             Estimate  Std. Error   t value
(Intercept)  3156.304    17.12144  184.3481
```

While the complete pooling approach ignores and likely underestimates the variation between groups, the no-pooling approach overstates it in modeling each group's specific mean separately and taking their average. Hence, making these methods subject to under or over-fitting the data if clusters are truly present, yet correlated to some extent. On the other hand, the partial-pooling approach, which is handled using multilevel modeling/analysis, is a compromise between the two extremes, using an average of pooled group means to represent the data as a way of adjusting for both cluster-specific variabilities and the variation between them. In this way, this type of model is a generalization of classical linear regression models, relaxing some of their strict assumptions, which when met, simply reduces to one of the aforementioned extreme cases. Evidently, when the data reflects either of these extremes, it may be preferred to fit a simple linear regression model, as it does not require the complexity of the multilevel approach to arrive at the same result. However, a partial-pooling estimate is not, in the same way, at risk for misrepresenting (potential groupings in) the data.

Calculating all three pooling mean estimates with respect to child birth weight, using both regression and their corresponding formulas, we notice that all three approaches produced nearly the same results (with a more noticeable difference in the formula-derived partial-pooling estimate). The reason behind this discrepancy (especially with regards to the different approaches used to compute the partial-pooling estimate) is unclear, although, it is possible for error to have played a role. Nonetheless, the regression estimates confirm that all three means are the same. This could be an indication that the small and consistent number of observations in each group, in tandem with the large number of groups present, results in group-level variability that is too small to yield an average of (unpooled or even partially pooled) mother-specific means that differs significantly (if at all) from the overall mean across individual birth weights in the data. Namely, the within-group variability is overpowered by the overall mean as the number of groups and corresponding samples is not large enough in comparison. This does not suggest however, that group-level variability is not present, but rather that it requires a more complex model (multilevel rather than classical linear regression) to be accurately represented.

## Question 4:

Suppose model M1:

$$Y_{ij} = \beta_0 + \beta_1 A_{i1} + \beta_2 A_{iD} + \epsilon_{ij}, \quad \text{where} \;\; \epsilon_{ij} \sim \text{N}(0, \sigma^2)$$

a. What type of model is M1? What are the main assumptions of this model?
b. Fit this model to the data. What are the estimates of the $\beta$ coefficients and $\sigma^2$ according to this model?
c. Interpret the regression coefficients.

**Solution**

a. This is a multiple linear regression model with predictors: mother's age at time of birth of her first child ($A_{i1}$) and age difference from that first birth ($A_{iD}$). Since it does not include group indicators (`mid`) or group-level predictors, this complete pooling model assumes that (the data comes from a population in which) there is no variation in birth weights across mothers. That is, it assumes a between-group variance of 0, and hence, gives an overall average of birth weights in the data ignoring mother-specific differences. Specifically, it models average birth weight as a function of mother age at first birth (centered at 21 years) and her age difference from this time. Additionally, as it is not a multilevel model (rather a classical linear regression model), it assumes, linearity between predictors and normally distributed mean outcome; independent observations; and equal residual variance among them (observations).

b. **M1**:

```
m1 <- lm(weight ~ age0 + ageDiff, data=birthwt)
display(m1)

             coef.est  coef.se
(Intercept)  3205.29     15.70
age0           31.61      2.45
ageDiff        13.21      2.80

Residual standard error: 558.5 on 4387 degrees of freedom
```

$$\beta_0 = 3205.29, \quad \beta_1 = 31.61, \quad \beta_2 = 13.21, \quad \sigma^2 = 558.5^2$$

c. **Coefficient Interpretation**:

- $\beta_0$: The average birth weight of a child, adjusting for mother's age at the time of her first birth and the difference in her age between her first and current births, is $3205.29 \approx 3205$ grams.
- $\beta_1$: The effect of a 1-unit increase in mother's age at the time of her first birth (centered at 21 years) on a child's birth weight, adjusting for the difference in her age form this time and the current child's birth, is $31.61 \approx 32$ grams. For example, if a mother was 20 at the time of her first child's birth (such that $A_{i1}$=20-21=-1), the expected birth weight of her $j$th child is $3205.29 - 31.61 = 3173.68 \approx 3174$ grams, adjusting for $A_{iD}$.
- $\beta_2$: The effect of a 1-unit increase in the difference in a mother's age since the time of her first birth on a child's birth weight, adjusting for this centered baseline age, is $13.21 \approx 13$ grams. For example, if a mother was 6 years older at the time of her $j$th child's birth than she was at the time of her first birth (such that $A_{iD}$=6), the $j$th child's expected birth weight is $3205.29 + (6 \times 13.21) = 3284.55 \approx 3285$ grams, adjusting for $A_{i1}$.
- $\sigma^2$: The variance of the residuals (errors) $\epsilon_{ij}$ is $558.5^2 = 311922.2 \approx 311922$. Namely, this value quantifies the variation of (individual) children birth weights around the overall mean. Specifically, it suggests that children birth weights deviate, on average, $558.5 \approx 559$ grams from the estimated mean birth weight across all children.

## Question 5:

Suppose model M2:

$$Y_{ij} = \beta_{0i} + \beta_1 A_{i1} + \beta_2 A_{iD} + \epsilon_{ij}, \quad \text{where } \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),\ \beta_{0i} \sim \mathrm{N}(\beta_0, \sigma^2_{\beta_0})$$

a. What type of model is M2? What are the main assumptions of this model?

b. Fit this model to the data.
c. How many $\beta$ coefficients are estimated according to this model?
d. Provide estimates and interpret the hyper-parameters of this regression model.

**Solution**

a. This is a multilevel linear regression model with predictors: mother's age at time of birth of her first child ($A_{i1}$) and age difference from that first birth ($A_{iD}$). Given that the intercept coefficient is defined as $\beta_{0i}$ as opposed to $\beta_0$ (such that $\beta_{0i} = \beta_0 + b_{0i}$), the model allows intercepts to vary across mothers, and hence, accounts for the group-level variation with respect to birth weight. Moreover, this partial pooling model adjusts for variability within groups and assumes a compromised variation in birth weights across mothers, modeling pooled mean estimates rather than overall or group-specific mean estimates. Specifically, it models average pooled birth weights as a function of mother age at first birth (centered at 21 years) and her age difference from this time. Additionally, in being a multilevel model, contrary to a classical linear model, it assumes neither equal residual variance among observations (as these are expected, but not required, to vary between groups); nor independence of observations, as those clustered within groups are expected to be dependent on one another.

b. **M2**:

```
m2 <- lmer(weight ~ age0 + ageDiff + (1|mid), data=birthwt)
display(m2)

             coef.est  coef.se
(Intercept)  3209.21     20.36
age0           31.63      3.94
ageDiff        12.19      2.30

Error terms:
  Groups    Name         Std.Dev.
  mid       (Intercept)  351.89
  Residual               433.93
```

c. **Estimated Coefficients**:

$$\beta_0 = 3209.21, \quad \beta_1 = 31.63, \quad \beta_2 = 12.19, \quad \sigma^2 = 433.93^2, \quad \sigma^2_{\beta_0} = 351.89^2$$

Aside from the hyper-parameters (interpreted in part (d) below), this model estimates a total of 3 $\beta$ coefficients in the fixed effects part of the model; $\beta_0$, $\beta_1$, and $\beta_2$. Additionally, the model fits a total of 878 random intercepts. That is, one $\beta_{0i}$ for each mother, $i$, in the data[2].

d. **Hyper-parameter Interpretation**:

- $\beta_0$: The fixed average birth weight of a child (assuming no random effect from the mother), adjusting for mother's age at the time of her first birth and the difference in her age between her first and current births, is $3209.21 \approx 3209$ grams.
- $\sigma^2$: The variance of the residuals (errors), $\epsilon_{ij}$, not accounted for by group-level variability (captured by the random intercept coefficient, $\beta_{0i}$), is $433.93^2 = 188295.2 \approx 188295$ grams. Namely, this value represents the variation of individual children birth weights around their mother-specific mean (within-group variability).
- $\sigma^2_{\beta_0}$: The variance of the random intercept errors, $b_{0i}$ (and equally, of $\beta_{0i}$, which has a mean of $\beta_0$), is $351.89^2 = 123826.6 \approx 123827$ grams. Namely, this value quantifies the variation of (individual) children birth weights around the overall mean. Specifically, it represents the between-group variation or the group-level variability around the mean of pooled mother-specific means.

[2]Each $\beta_{0i}$, which is the sum of the fixed intercept, $\beta_0$, and the random effect of the $i^{\text{th}}$ mother, $b_{0i}$.

## Question 6:

Compare the respective $\beta$ coefficients between M1 and M2. Which model do you think best fits the data? Explain.

**Solution**

From the (fixed) $\beta$ coefficients estimated in both models, we see that allowing intercepts to vary across mothers (M2), didn't have a significant impact on fixed effects. Particularly, (referencing the examples used previously), in M2 we have:

- $\beta_0$: The fixed average birth weight of a child (assuming no random effect from the mother), adjusting for mother's age at the time of her first birth and the difference in her age between her first and current births, is $3209.21 \approx 3209$ grams, which is only about 4 grams greater what the estimate from M1.
- $\beta_1$: The effect of a 1-unit increase in mother's age at the time of her first birth (centered at 21 years) on a child's birth weight, adjusting for the difference in her age form this time and the current child's birth, is $31.63 \approx 32$ grams. So, if a mother was 20 at the time of her first child's birth (such that $A_{i1}$=20-21=-1), the expected birth weight of her $j$th child according to this model would be $3209.21 - 31.63 = 3177.58 \approx 3178$ grams, adjusting for $A_{iD}$, which is also approximately 4 grams greater than what is estimated by M1.
- $\beta_2$: The effect of a 1-unit increase in the difference in a mother's age since the time of her first birth on a child's birth weight, adjusting for this centered baseline age, is $12.19 \approx 12$ grams. So, if a mother was 6 years older at the time of her $j$th child's birth compared to her first birth (such that $A_{iD}$=6), the $j$th child's expected birth weight is $3209.21 + (6 \times 12.19) = 3282.35 \approx 3282$ grams, adjusting for $A_{i1}$. When compared to M1, which would predict this outcome to be 3285, we notice an increase of merely 3 grams.

In addition to the small change in fixed coefficient estimates (and hence adjusted predicted birth weights), the change in residual standard error from 558.5 in M1 to 443.93 in M2, suggests that randomizing intercepts didn't account for a significant amount of individual level variation, despite showing a comparable difference between groups. Moreover, calculating the respective Intraclass Correlation Coefficient (ICC), using `ICC(outcome="weight", group="mid", data=birthwt)`, which gave a value of $\approx 0.4175$, further validates our speculation that variability between clusters is fairly small, yet significant enough to choose the multilevel model (M2) over the classical linear regression model (M1).

## Question 7:

Do you believe that the order of birth is strongly associated with birth weight after adjusting for $A_{i1}$ and $A_{iD}$? To answer this question:

a. Fit an appropriate model, M3, to the data.
b. Use appropriate notation to write the regression equation for M3.
c. Interpret the regression coefficient(s) for the order of birth from the fixed-effects part of M3.

*Note: This is an open-ended question. There is no restriction on the number or type of parameters you can assume for this model (e.g., random intercepts, and or slopes, etc.), as long as the model is meaningful; can address this particular question; and there are no convergence issues when fitting it to the data.

**Solution**

a. **M3**:

```
m3 <- lmer(weight ~ age0 + ageDiff + order + (1|mid), data=birthwt)
display(m3)

             coef.est  coef.se
(Intercept)  3207.93     25.39
age0           31.64      3.95
ageDiff        11.71      6.15
order           1.05     12.38

Error terms:
  Groups    Name         Std.Dev.
  mid       (Intercept)  351.95
  Residual               433.98
```

b. **Regression Equation**:

$$Y_{ij} = \beta_{0i} + \beta_1 A_{i1} + \beta_2 A_{iD} + \beta_3 O_{ij} + \epsilon_{ij}, \quad \text{where } \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),\ \beta_{0i} \sim \mathrm{N}(\beta_0, \sigma^2_{\beta_0})$$
$$\beta_0 = 3207.93, \quad \beta_1 = 31.64, \quad \beta_2 = 11.71, \quad \beta_3 = 1.05$$
$$\sigma^2 = 433.98^2, \quad \sigma^2_{\beta_0} = 351.95^2$$

c. **Coefficient Interpretation**: $\beta_3 O_{ij}$

- $\beta_3$: The effect of a 1-unit increase in child birth order on the $j^{\text{th}}$ child's birth weight (individual level predictor), adjusting for a mother's age at the time of her first birth and the difference in her age form this time and the current child's birth, is $1.05 \approx 1$ gram. Thus, according to this model, a child's birth order is roughly proportional to the number of grams added to their expected birth weight, adjusting for other model covariates. Hence, for example, the model would estimate that the $j^{\text{th}}$ mother's second child is approximately 1 gram heavier than her first, adjusting for effects of $A_{i1}$ and $A_{iD}$.

Not only is this effect comparatively small, but performing a likelihood ratio test (LRT) between M2 and M3 (assuming that M3 is nested in M2, as the inclusion of this predictor is the only change made to the multilevel model), further validates the insignificance of birth order in predicting a child's birth weight. That is, with a p-value of $\approx 0.9333 \geq 0.05$, we fail to reject the null hypothesis that both model's fit the data equally well. For this reason, in tandem with the fact that this model produced, beta coefficients, standard errors, and AIC values that were nearly identical to those given by M2, it follows that a child's order of birth is not associated with birth weight (after adjusting for $A_{i1}$ and $A_{iD}$).

## Question 8:

Low birth weight (LBW), namely birth weights less than 2500 grams, is associated with increased risk of health complications for the child. Using an appropriate model, investigate how the order of birth is associated with LBW after adjusting for $A_{i1}$ and $A_{iD}$. To answer this question:

a. Fit an appropriate model, M4, to the data.
b. Write the regression equation using appropriate notation.
c. Provide estimates for the regression coefficients.
d. Interpret the beta coefficient(s) for the effect of the order of birth on the risk of LBW.

*Note: This is an open-ended question. There is no restriction on the number or type of parameters you can assume for this model (e.g., random intercepts, and or slopes, etc.), as long as the model is meaningful; can address this particular question; and there are no convergence issues when fitting it to the data.

**Solution**

a. **M4**:

```
m4 <- glmer(LBW ~ age0 + ageDiff + order + (1|mid),
            family=binomial(link="logit"), data=birthwt)
```

```
             coef.est  coef.se
(Intercept)     -3.08     0.20
age0            -0.09     0.02
ageDiff          0.01     0.04
order           -0.03     0.09

Error terms:
  Groups    Name         Std.Dev.
  mid       (Intercept)     1.33
  Residual                  1.00
```

b. **Regression Equation**:

$$\text{logit}(\text{E}[Y_{ij}]) = \beta_{0i} + \beta_1 A_{i1} + \beta_2 A_{iD} + \beta_3 O_{ij} + \epsilon_{ij}, \quad \text{where} \;\; \epsilon_{ij} \sim \text{N}(0, \sigma^2), \;\; \beta_{0i} \sim \text{N}(\beta_0, \sigma^2_{\beta_0})$$

Where:

- $Y_{ij}$ is the predicted outcome of having low birth weight (LBW) (yes=1, or no=0) and $\text{E}[Y_{ij}] = \mu_Y$ is the risk of LBW ($P(Y_{ij} = 1)$) for the $i^{\text{th}}$ mother's $j^{\text{th}}$ child, such that,
$$\text{logit}(\mu_Y) = \log\left(\frac{\mu_Y}{1-\mu_Y}\right) = \log\Big[\text{odds}\big[P(Y_{ij} = 1)\big]\Big];$$
- $\beta_{0i}$ is the random (mother-specific) intercept, such that $\beta_{0i} = \beta_0 + b_{0i}$, where $\beta_0$ is the log-odds of having LBW, adjusting for model covariates, and $b_{0i}$ is the (random) effect of being born to the $i^{\text{th}}$ mother on the log-odds that $Y_{ij} = 1$;
- $\beta_1$ is fixed slope for $A_{i1}$, which gives the effect of a 1-unit increase in a mother's age at the time of her first child's birth on the log-odds that $Y_{ij} = 1$, adjusting for other model covariates;
- $\beta_2$ is fixed slope for $A_{iD}$, which gives the effect of a 1-unit increase in difference in mother's age from her first and $j^{\text{th}}$ births on the log-odds that her $j^{\text{th}}$ child has LBW, adjusting for other model covariates; and
- $\beta_3$ is fixed slope for $O_{ij}$, which gives the effect of a 1-unit increase in the child's birth order, $j$, on the log-odds that $Y_{ij} = 1$, adjusting for other model covariates.

c. **Coefficient Estimates**:

$$\beta_0 = -3.08, \quad \beta_1 = -0.09, \quad \beta_2 = 0.01, \quad \beta_3 = -0.03$$

$$\sigma^2 = 1, \quad \sigma^2_{\beta_0} = 1.33^2$$

d. **Coefficient Interpretation**: $\beta_3 O_{ij}$

- $\beta_3$: The effect of a 1-unit increase in the child's birth order, $j$, on the log-odds that $Y_{ij} = 1$ (having LBW), adjusting for a mother's age at the time of her first birth and the difference in her age form this time and the current child's birth, is $-0.03$. That is, according to this model, each unit increase in birth order multiplies the odds of having LBW by a factor of $e^{\beta_3} = 0.9704455$, adjusting for the effects of $A_{i1}$ and $A_{iD}$.

As with M3, this model suggests that order is not significant in estimating a child's risk of LBW.

## Code

```
## Importing Data

birthwt <- read.csv("/Users/antonellabasso/Desktop/PHP2517/DATA/birthwt.csv")
birthwt
```

```
## Descriptive Statistics

str(birthwt)
nrow(birthwt) # 4,390 observations/individual children born to
length(unique(birthwt$mid)) # 878 mothers
# 878 children in each order group (first borns, second borns,...)

# missing data
birthwt_na <- apply(birthwt, 2, function(x) sum(is.na(x)))
birthwt_na[birthwt_na != 0] # no missing values of any kind

# general info
mean(birthwt$weight) # mean birth weight = 3156
range(birthwt$weight) # birth weight range: 312-5528
mean(birthwt$age) # mean mother's age = 21
range(birthwt$age) # mother's ages range: 12-42
mean(birthwt[which(birthwt$order==1),]$age) # mean mother's age at first birth = 18
mean(birthwt[which(birthwt$order==5),]$age) # mean mother's age at last birth = 25
range(birthwt[which(birthwt$order==5),]$ageDiff) # diff. in mother's age from first-last born range: 3-
mean(birthwt[which(birthwt$order==5),]$ageDiff) # mean diff. in mother's age from first-last born = 7

# variable summary
CreateTableOne(data=birthwt)

# mode function
getmode <- function(x) {
  unique_x <- unique(x)
  unique_x[which.max(tabulate(match(x, unique_x)))]
}

getmode(birthwt$age) # mode mother's age = 18
getmode(birthwt[which(birthwt$order==5),]$age) # mode mother's age at last birth = 23
getmode(birthwt[which(birthwt$order==1),]$age) # mode mother's age at first birth = 16

# new df with mother id (mid) changed to 1-878
birthwt2 <- birthwt %>%
  mutate(mid=sort(rep(1:878, 5)))
birthwt2 <- birthwt2[, c(2, 1, 6, 3, 4, 5, 7, 8)]

# descriptive statistics wrt birth weight by mother (group)
ds_mother <- birthwt2 %>%
  group_by(mid) %>%
  summarise(children=n(), # count, ni
            mean_weight=mean(weight), # mean birth weight, Yi
            var_weight=var(weight), # var birth weight, s2i
            mean_age=mean(age),
            var_age=var(age))
```

```r
# descriptive statistics wrt birth weight by child order
ds_order <- birthwt %>%
  group_by(order) %>%
  summarise(children=n(), # count
            mean_weight=mean(weight), # mean birth weight
            var_weight=var(weight), # var birth weight
            mean_age=mean(age),
            var_age=var(age),
            mean_diff=mean(ageDiff),
            var_diff=var(ageDiff))

# factorizing categorical variables
#birthwt2$order <- as.factor(birthwt2$order) # group (order)
#birthwt2$mid <- as.factor(birthwt2$mid) # mother id
#birthwt2$cid <- as.factor(birthwt2$cid) # child id
```

```r
## EDA Table

# descriptive stats of birth weight (response) and other predictors by mother (i=1, 2,...,5)
dstats1 <- ds_mother %>%
  mutate(mean_weight=round(mean_weight), var_weight=round(var_weight),
         mean_age=round(mean_age), var_age=round(var_age)) %>%
  rename("Mother Index"=mid,
         "Children"=children,
         "Birth Weight Mean"=mean_weight,
         "Birth Weight Var"=var_weight,
         "Mother's Age Mean"=mean_age,
         "Mother's Age Var"=var_age)

# descriptive stats of birth weight (response) and other predictors by child order
dstats2 <- ds_order %>%
  mutate(mean_weight=round(mean_weight), var_weight=round(var_weight),
         mean_age=round(mean_age), var_age=round(var_age),
         mean_diff=round(mean_diff), var_diff=round(var_diff)) %>%
  rename("Child Index"=order,
         "Children"=children,
         "Birth Weight Mean"=mean_weight,
         "Birth Weight Var"=var_weight,
         "Mother's Age Mean"=mean_age,
         "Mother's Age Var"=var_age,
         "Age Diff. Mean"=mean_diff,
         "Age Diff. Var"=var_diff)

knitr::kable(dstats1[1:5,],
             caption="Descriptive Statistics for First 5 Mothers",
             digits=4, booktabs=T, linesep="") %>%
  add_header_above(c(" "=1, "Primary Outcome"=3, "Other Predictors"=2)) %>%
  kable_styling(latex_options=c("hold_position", "scale_down"))

knitr::kable(dstats2,
             caption="Descriptive Statistics by Child Order",
             digits=4, booktabs=T, linesep="") %>%
  add_header_above(c(" "=1, "Primary Outcome"=3, "Other Predictors"=4)) %>%
```

```
  kable_styling(latex_options=c("hold_position", "scale_down"))
```

```
## Density Plots (Histograms)

# Birth Weight Density
hist(birthwt$weight,
     probability=TRUE, col="lavender", breaks=20,
     main="Figue 1: Birth Weight Density",
     xlab="Birth Weight (g)")
lines(density(birthwt$weight), col="red")
```

```
# Centered Mother's Age Density
hist(birthwt$age_c,
     probability=TRUE, col="thistle2", breaks=20,
     main="Figue 2: Centered Mother's Age Density",
     xlab="Mother's Age, Centered at 21")
lines(density(birthwt$age_c), col="red")

# Mother's Age Difference (Between Child 1-5) Density
hist(birthwt[which(birthwt$order==5),]$ageDiff,
     probability=TRUE, col="thistle", breaks=10,
     main="Figure 3: Mother's Age Difference Density (Between Child 1-5)",
     xlab="Mother's Age Difference (Between Child 1-5)")
lines(density(birthwt[which(birthwt$order==5),]$ageDiff), col="red")
```

```
# Densities of Birth Weight by Child Order
par(mfrow=c(2, 2))
colnames <- seq(1, 5, 1)
for (i in colnames) {
    hist(birthwt[which(birthwt$order==i), "weight"],
         probability=TRUE, col="lightblue",
         #xlim=c(0, 17), breaks=seq(0, 17, 1),
         main=paste0("Child ", i), xlab="Birth Weight (g)")
    dens <- density(birthwt[which(birthwt$order==i), "weight"])
    lines(dens, col="red")
}
```

```
## EDA Plots

# Spread of Birth Weight by Child Order
p1 <- ggplot(birthwt, aes(x=as.factor(order), y=weight, color=as.factor(order))) +
  geom_boxplot() +
  labs(title="Figure 5: Spread of Birth Weight by Child Order",
       x="Child",
       y="Birth Weight (g)",
       color="")

# Birth Weight | Mother's Age by Child Order
p2 <- ggplot(birthwt, aes(x=age, y=weight)) +
  geom_point(aes(color=as.factor(order))) +
  geom_smooth(method="loess", se=FALSE) +
  labs(title="Figure 6: Birth Weight vs. Mother's Age by Child Order",
         x="Mother's Age (years)",
         y="Birth Weight (g)",
         color="Child")
```

```r
mean_bw_fc <- birthwt %>%
  group_by(mid) %>%
  summarise(mean_bw=mean(weight),
            mean_age=mean(age),
            age0=age0)

# # Mean Birth Weights | Mother's Age at First Child's Birth
# p3 <- ggplot(mean_bw_fc, aes(x=age0, y=mean_bw)) +
#   geom_point(color="seagreen") +
#   geom_smooth(method="loess", se=FALSE) +
#   labs(title="Figure 7: Mean Birth Weights by Mother's Age at First Child's Birth",
#        x="Mother's Age at First Child's Birth (years)",
#        y="Mean Birth Weight (g)")

# Mean Birth Weights | Mother's Mean Age
p4 <- ggplot(mean_bw_fc, aes(x=mean_age, y=mean_bw)) +
  geom_point(color="seagreen") +
  geom_smooth(method="loess", se=FALSE) +
  labs(title="Figure 8: Mean Birth Weights by Mean Mother's Age",
       x="Mother's Mean Age (years)",
       y="Mean Birth Weight (g)")
```

```r
## Trajectories
set.seed(47)

mid_5 <- sample(unique(birthwt2$mid), 5, replace=FALSE)

traj <- ggplot(birthwt2[which(birthwt2$mid %in% mid_5),], aes(x=order, y=weight, color=as.factor(mid)))
  geom_point() +
  geom_line() +
  labs(title="Figure 9: Birth Weight Trajectories for Randomly Selected Mothers",
       x="Child",
       y="Birth Weight (g)",
       color="Mother Index")
```

```r
# factorizing variables
birthwt$mid <- as.factor(birthwt$mid)
birthwt2$mid <- as.factor(birthwt2$mid)
birthwt$cid <- as.factor(birthwt$cid)
birthwt2$cid <- as.factor(birthwt2$cid)
birthwt$order <- as.factor(birthwt$order)
birthwt2$order <- as.factor(birthwt2$order)
```

```r
## Pooling Estimates From Formula

# complete pooling
mu_all <- mean(birthwt2$weight)
var_all <- var(birthwt2$weight)

# no-pooling
group_means <- c()
for (i in 1:length(unique(birthwt2$mid))){
  mean_i <- mean(birthwt2[which(birthwt2$mid==i),]$weight)
  group_means <- c(group_means, mean_i)
```

```
}
group_vars <- c()
for (i in 1:length(unique(birthwt2$mid))){
  var_i <- var(birthwt2[which(birthwt2$mid==i),]$weight)
  group_vars <- c(group_vars, var_i)
}

# partial-pooling
pooled_means <- c()
for (i in 1:length(unique(birthwt2$mid))){
  num <- ((5*group_means[i])/group_vars[i])+(mu_all/var_all)
  denom <- (5/(group_vars[i]))+(1/var_all)
  a_i <- num/denom
  pooled_means <- c(pooled_means, a_i)
}

c(mu_all, mean(group_means), mean(pooled_means)) # mean estimates
```

```
## Pooling Estimates From Regression

# complete pooling
m_cp <- lm(weight~ 1, data=birthwt) # 3156.304

# no-pooling
m_np <- lm(weight~ -1+mid, data=birthwt)
m_np2 <- lm(m_np$coef~ 1) # mean of means = 3156.304

# partial-pooling
m_pp <- lmer(weight~ 1+(1|mid), data=birthwt) # 3156.304

# mean estimates
m_cp$coef
m_np2$coef
summary(m_pp)$coef
```

```
## M1

m1 <- lm(weight ~ age0 + ageDiff, data=birthwt) # complete pooling

summary(m1)
display(m1)
coef(m1)
```

```
## M2

m2 <- lmer(weight ~ age0 + ageDiff + (1|mid), data=birthwt) # partial-pooling

summary(m2)
display(m2)
```

```
## ICC (assessing the level of between-cluster variability)

ICC(outcome="weight", group="mid", data=birthwt) # 0.4175
```

```
## M3

m3 <- lmer(weight ~ age0 + ageDiff + as.numeric(order) + (1|mid), data=birthwt)

summary(m3)
display(m3)
```

```
## LRT (between M2 and M3)

anova(m2, m3) # p-value ~ 0.9 > 0.05
```

```
## Adding New Binary LBW Variable to Data

birthwt$LBW <- rep(0, nrow(birthwt))

for (i in 1:nrow(birthwt)){
  if (birthwt$weight[i]<2500){
    birthwt$LBW[i] <- 1
  }
}

birthwt$LBW <- as.factor(birthwt$LBW)
```

```
## M4

m4 <- glmer(LBW ~ age0 + ageDiff + as.numeric(order) + (1|mid),
            family=binomial(link="logit"), data=birthwt)

summary(m4)
display(m4)
```