

# Georgia Birth Weight Modeling

Alejandro Hernandez

2024-06-04

```
birthwt <- read.csv("data/cdc birthwt.csv")

names(birthwt) <- c("mother.id", "birth.order", "birth.weight", "maternal.age", "child.id")

birthwt <- birthwt %>%
  # create birth weight binary factor
  mutate(birth.weight.binary = ifelse(birth.weight < 2500, 1, 0)) %>%
  # create a interpregnancy variable
  arrange(mother.id, birth.order) %>%
  group_by(mother.id) %>%
  mutate(interval = maternal.age - lag(maternal.age)) %>%
  ungroup

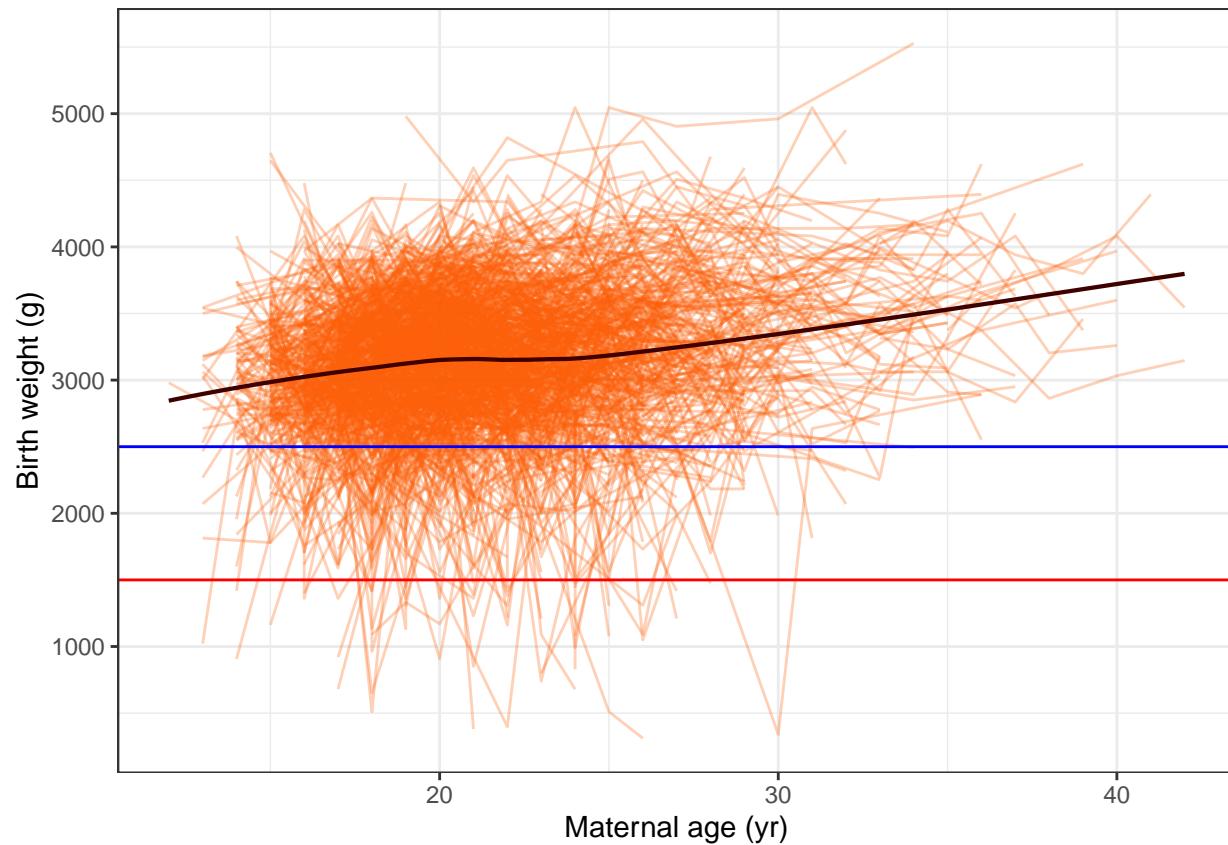
birthwt %>% filter(interval < 0)

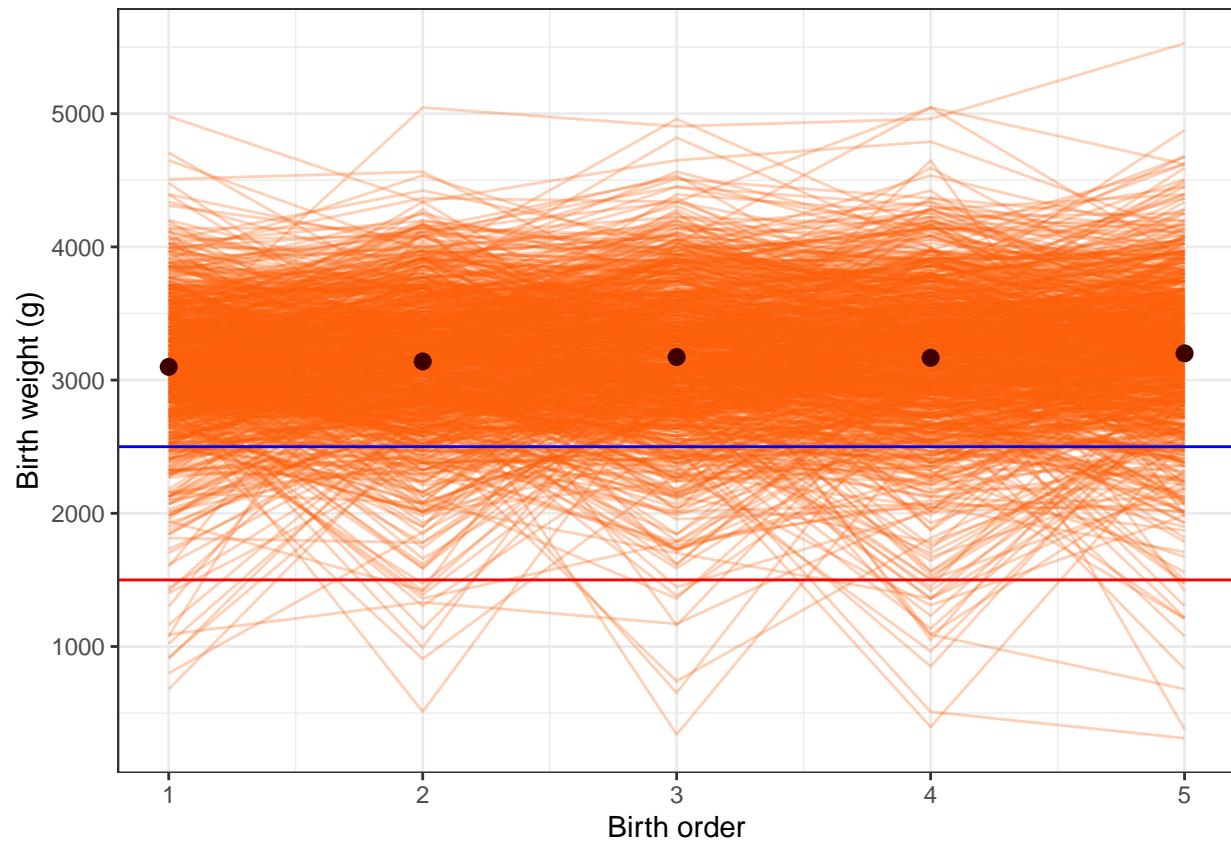
## # A tibble: 2 x 7
##   mother.id birth.order birth.weight maternal.age child.id birth.weight.binary
##       <int>      <int>      <int>      <int>      <int>              <dbl>
## 1     57939          3        3430         19      1063                 0
## 2    212988          3        2470         20      3648                 1
## # i 1 more variable: interval <int>

# mothers 57939 and 212988 both have negative intervals for their third births
# we will remove all their data and do not expect measurable changes in future
# analysis, their data is not special among the sample
birthwt <- birthwt %>% filter(!mother.id %in% c(57939, 212988))
```

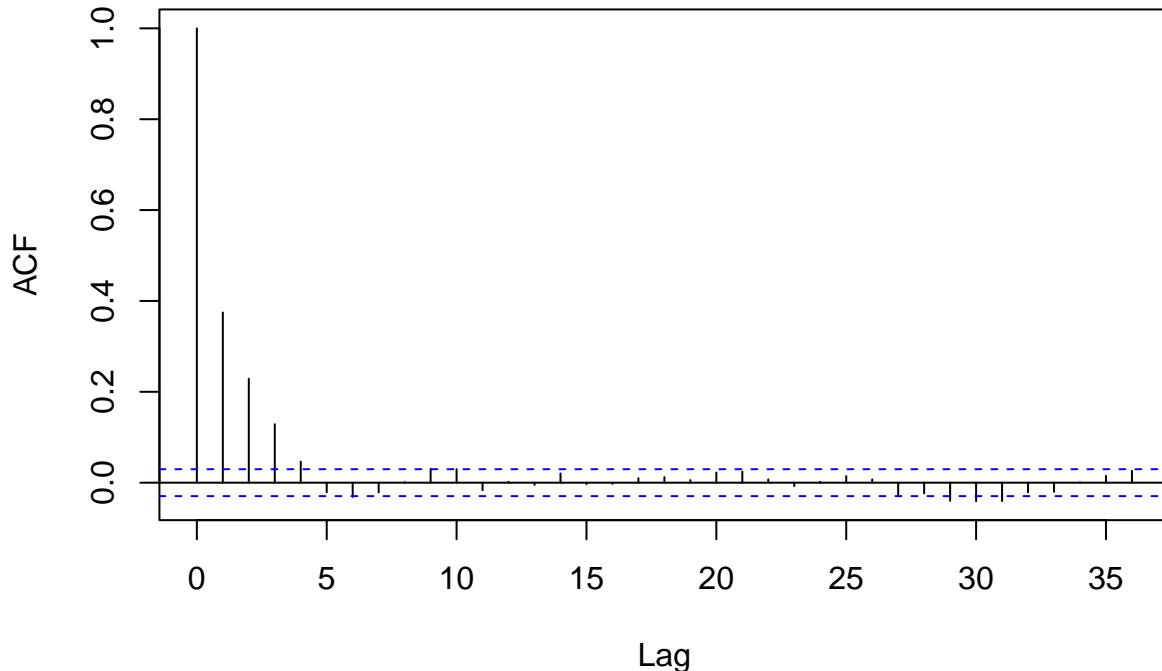
## Modeling correlation

```
## `geom_smooth()` using formula = 'y ~ x'
```





## Series birthwt\$birth.weight



We presume that average population effects on birth weight will be well-modeled by a generalized estimating equation (GEE) model with a auto-regressive correlation structure, based on the decaying pattern of the ACF graph above.

For the same reason, we expect the individual-specific effects on birth weight to be well-modeled by a linear mixed-effects model (LMM) that also assumes an auto-regressive correlation structure.

We elect to fit many LMM and GEE models, assuming various correlation structures. We expect to report coefficients and performance diagnostics for four models:

- (1) LMM with no covariates, random intercept and fixed slope, and no assumed correlation structure.
- (3) LMM with covariate for birth order, random intercept and fixed slope, and assumed auto-regressive correlation structure.
- (4) GEE with no predictors and auto-regressive working correlation structure.
- (5) GEE with covariate for birth order and auto-regressive working correlation structure.

### Modeling birth weight

```
#### Linear Mixed-Effects Models ####
# Random intercept, fixed slope; No covariates
lmm0 <- nlme::lme(birth.weight ~ maternal.age,
                    random = ~ 1 | mother.id,
                    data = birthwt)
```

```

# Random intercept, fixed slope; with birth order covariate
lmm1 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ 1 | mother.id,
             data = birthwt)

# Random intercept and slope; with birth order covariate
lmm2 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ maternal.age | mother.id,
             data = birthwt)

# Random intercept and fixed slope; with birth order covariate and auto-regressive correlation structure
lmm3 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ 1 | mother.id, correlation = corAR1(),
             data = birthwt)

# Random intercept and slope; with birth order covariate and auto-regressive correlation structure
# WILL NOT COMPLETE; DOES NOT CONVERGE
lmm4 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ maternal.age | mother.id, correlation = corAR1(),
             data = birthwt)

## Error in lme.formula(birth.weight ~ maternal.age + birth.order, random = ~maternal.age | : nlminb pr
##   message = iteration limit reached without convergence (10)

# Random intercept and slope; with birth order covariate and exchangeable
# correlation structure
lmm5 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ maternal.age | mother.id, correlation = corCompSymm(),
             data = birthwt)

# Random intercept and fixed slope; with birth order covariate and exponential
# correlation structure
lmm6 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ 1 | mother.id, correlation = corExp(),
             data = birthwt)

# Random intercept and slope; with birth order covariate and exponential
# correlation structure
# WILL NOT COMPLETE; DOES NOT CONVERGE
lmm7 <- lme(birth.weight ~ maternal.age + birth.order,
             random = ~ maternal.age | mother.id, correlation = corExp(),
             data = birthwt)

## Error in lme.formula(birth.weight ~ maternal.age + birth.order, random = ~maternal.age | : nlminb pr
##   message = iteration limit reached without convergence (10)

##### Generalized Estimating Equations #####
# Auto-regressive correlation structures assume that the correlation between
# observations decreases as the time between them increases
gee0 <- geepack::geeglm(birth.weight ~ maternal.age, id = mother.id,
                        corstr = "ar1", data = birthwt)

```

```

gee1 <- geeglm(birth.weight ~ maternal.age + birth.order, id = mother.id,
                 corstr = "ar1", data = birthwt)

# Exchangeable correlation structures assume that all pairs of observations
# within a subject have the same correlation, regardless of the time interval
# between them
gee2 <- geeglm(birth.weight ~ maternal.age, id = mother.id,
                 corstr = "exchangeable", data = birthwt)

gee3 <- geeglm(birth.weight ~ maternal.age + birth.order, id = mother.id,
                 corstr = "exchangeable", data = birthwt)

# Independent correlation structures assume that all measurements within a
# subject are uncorrelated
gee4 <- geeglm(birth.weight ~ maternal.age, id = mother.id,
                 corstr = "independence", data = birthwt)

gee5 <- geeglm(birth.weight ~ maternal.age + birth.order, id = mother.id,
                 corstr = "independence", data = birthwt)

```

## Diagnostics

```

## [1] "AIC scores, a measure of relative goodness of fit"

## Warning in AIC.default(lmm0, lmm1, lmm2, lmm3, lmm5, lmm6): models are not all
## fitted to the same number of observations

##      df      AIC
## lmm2  7 66854.91
## lmm5  8 66856.91
## lmm6  6 66870.87
## lmm3  6 66870.87
## lmm1  5 66901.08
## lmm0  4 66915.16

## [1] "Variance/correlation of LMM Model 0"

## mother.id = pdLogChol(1)
##          Variance StdDev
## (Intercept) 125784.5 354.6611
## Residual    188559.7 434.2346

## [1] "Variance/correlation of LMM Model 1"

## mother.id = pdLogChol(1)
##          Variance StdDev
## (Intercept) 123782.0 351.8267
## Residual    188641.4 434.3287

## [1] "Variance/correlation of LMM Model 2"

```

```

## mother.id = pdLogChol(maternal.age)
##          Variance StdDev   Corr
## (Intercept) 178298.6987 422.25431 (Intr)
## maternal.age    597.4044  24.44186 -0.752
## Residual      182165.8101 426.80887

## [1] "Variance/correlation of LMM Model 3"

## mother.id = pdLogChol(1)
##          Variance StdDev
## (Intercept) 110914.3 333.0379
## Residual     200915.8 448.2364

## [1] "Variance/correlation of LMM Model 5"

## mother.id = pdLogChol(maternal.age)
##          Variance StdDev   Corr
## (Intercept) 188305.2602 433.94154 (Intr)
## maternal.age    597.4176  24.44213 -0.732
## Residual      172159.0893 414.92058

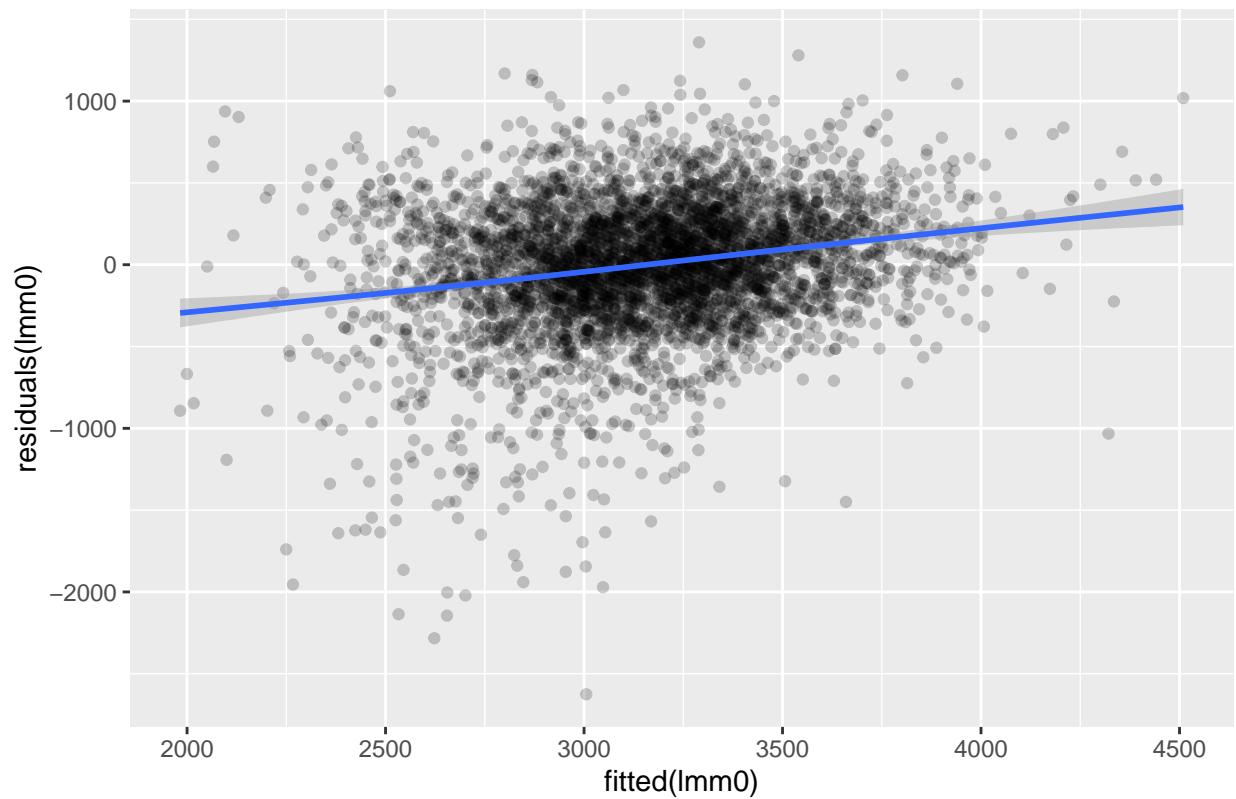
## [1] "Variance/correlation of LMM Model 6"

## mother.id = pdLogChol(1)
##          Variance StdDev
## (Intercept) 110915.0 333.0391
## Residual     200915.7 448.2362

## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

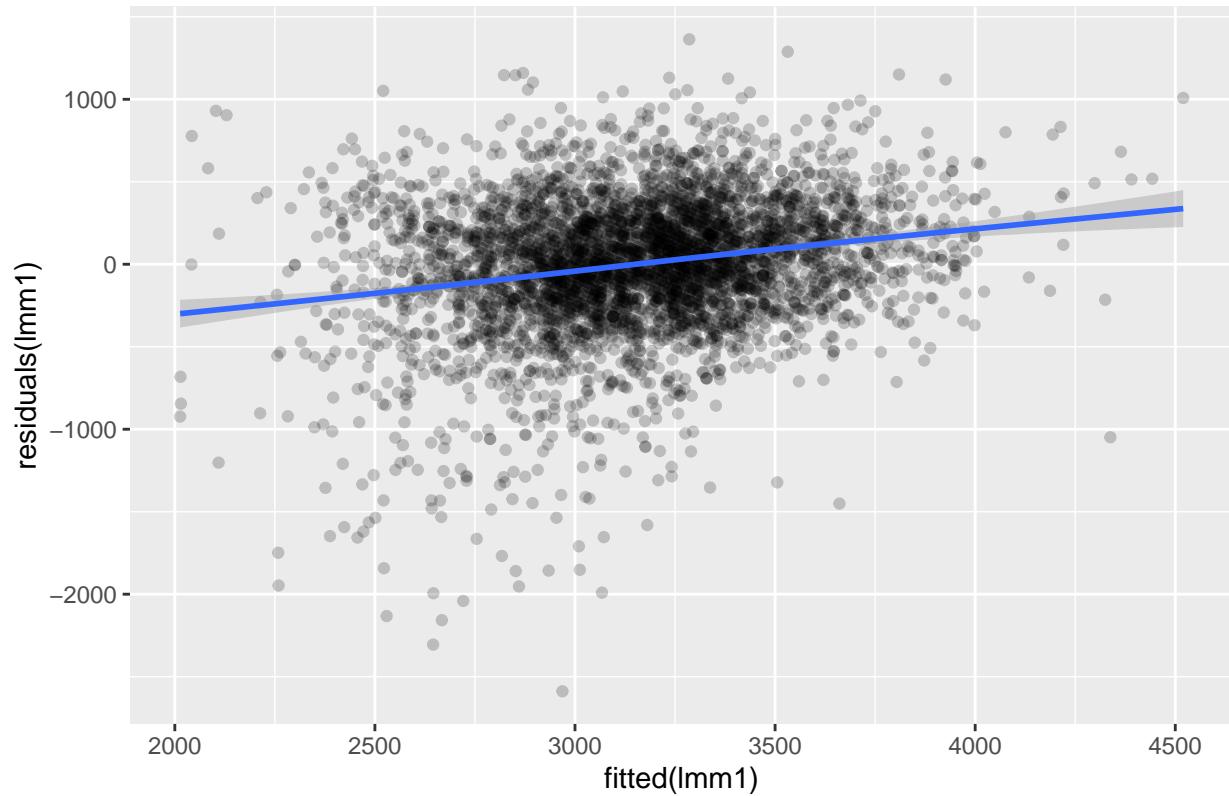
```

## LMM 0



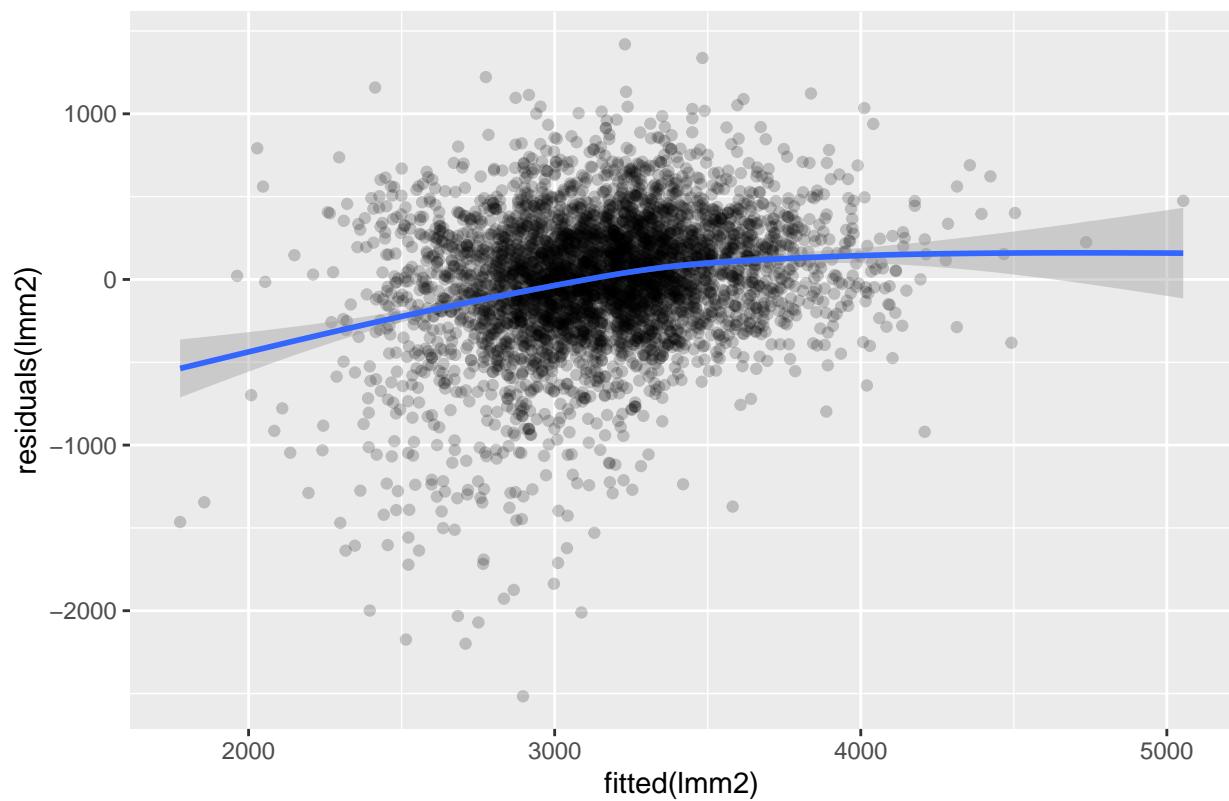
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## LMM 1



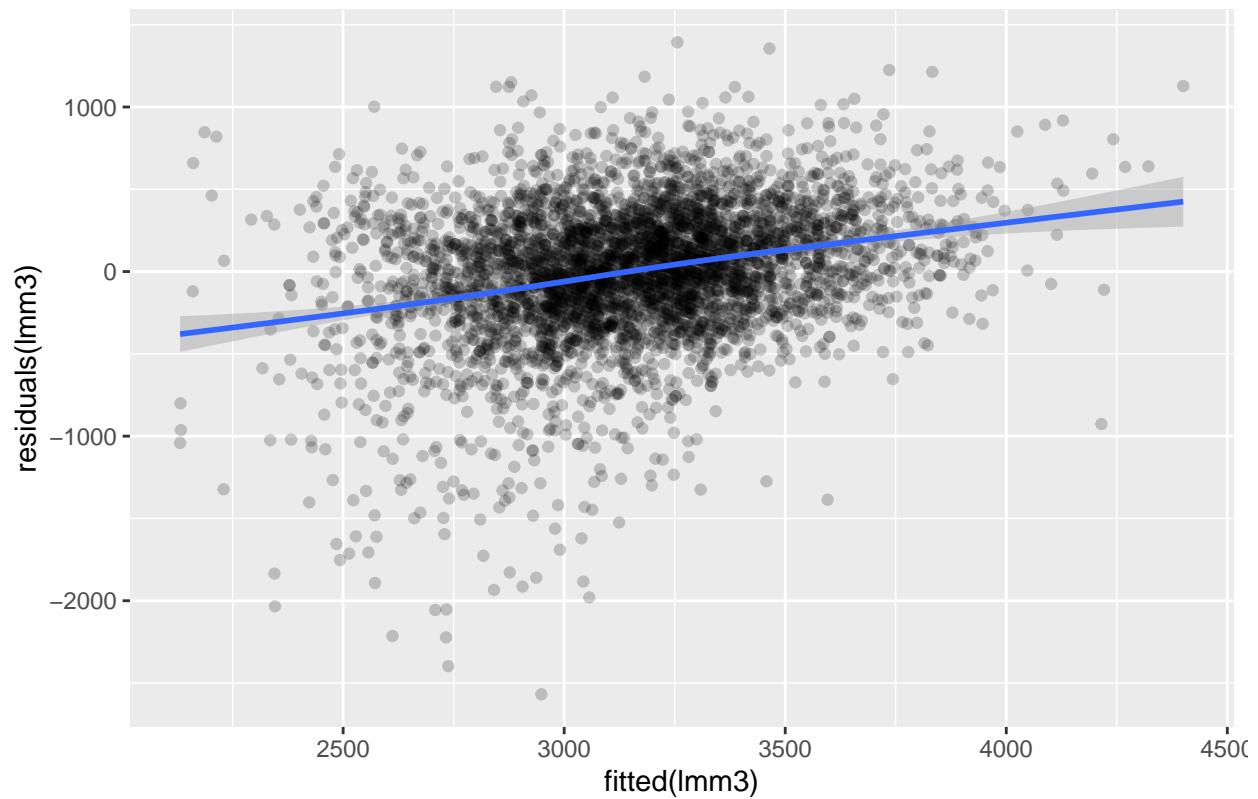
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## LMM 2



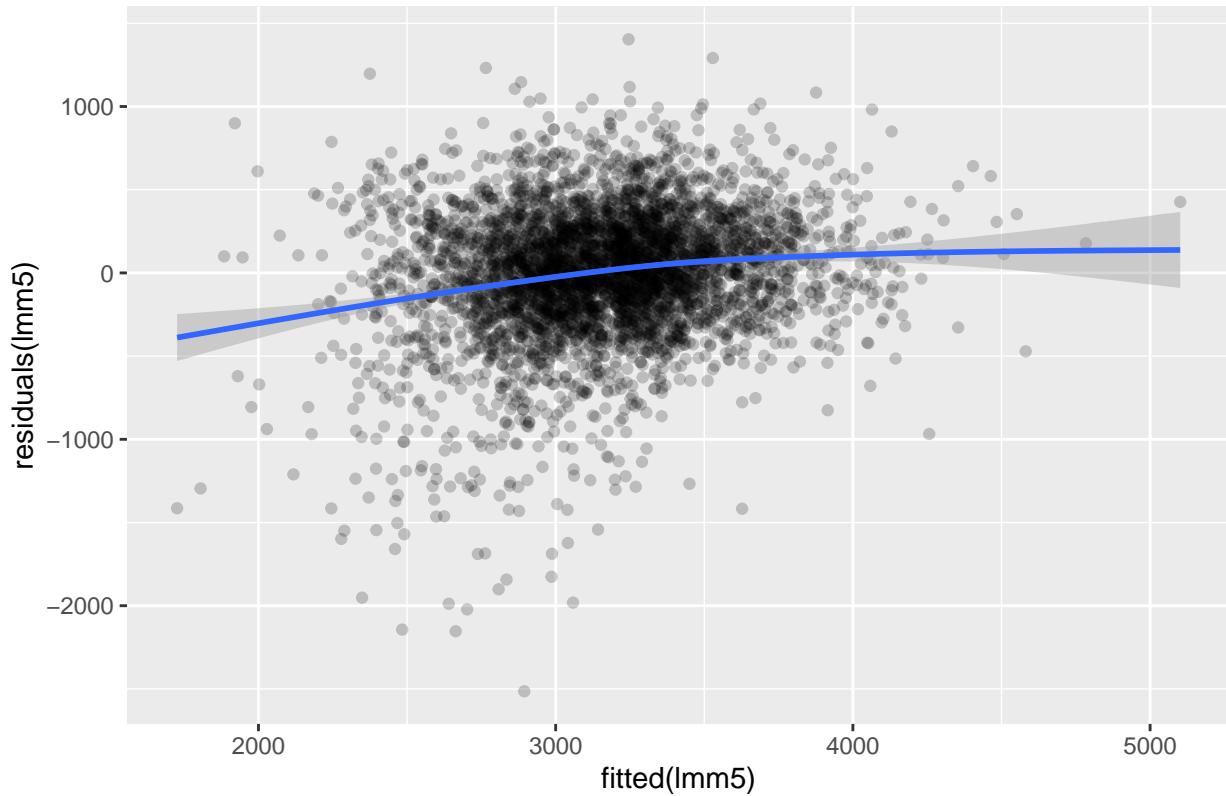
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

### LMM 3



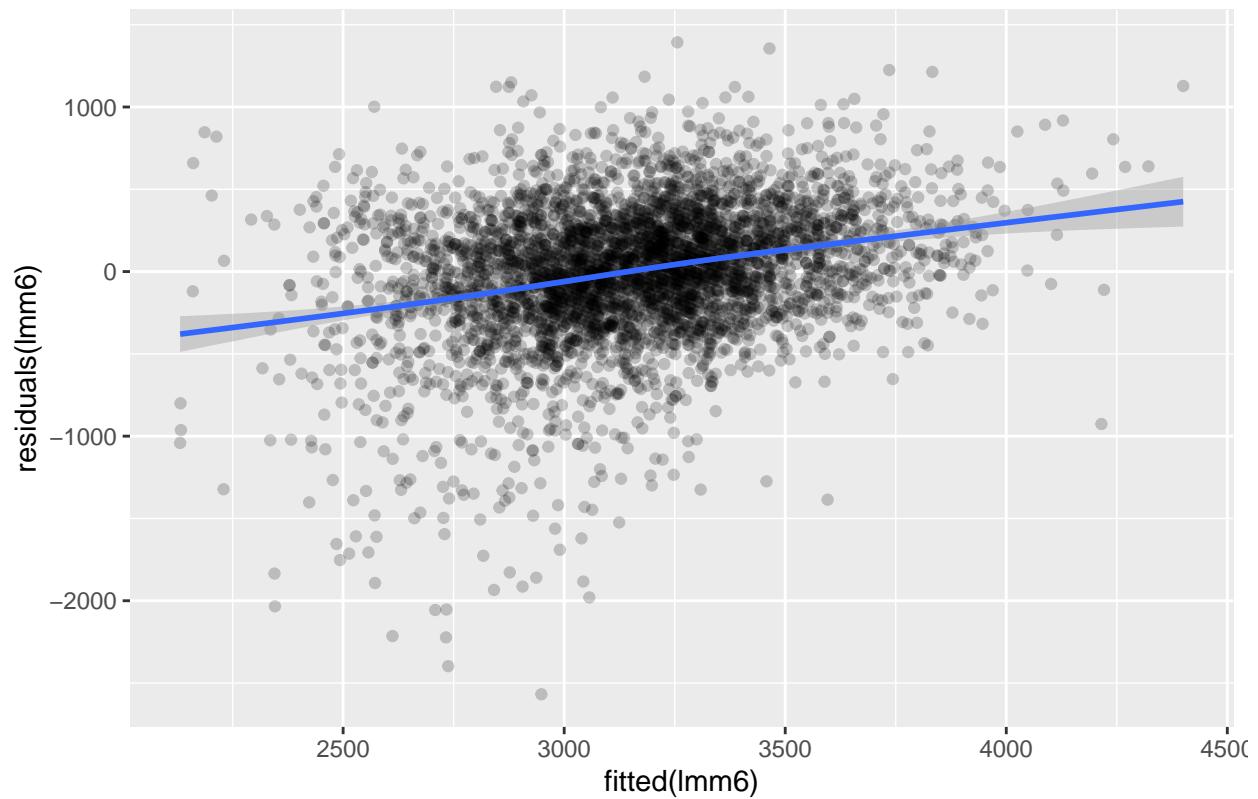
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## LMM 5

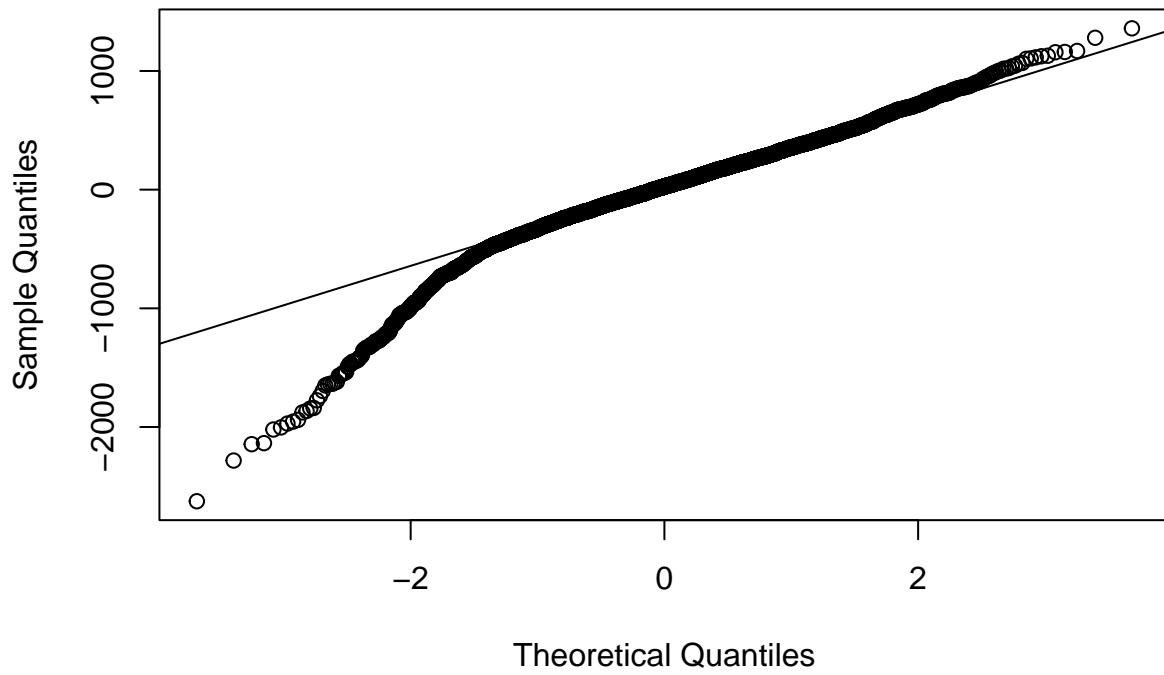


```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

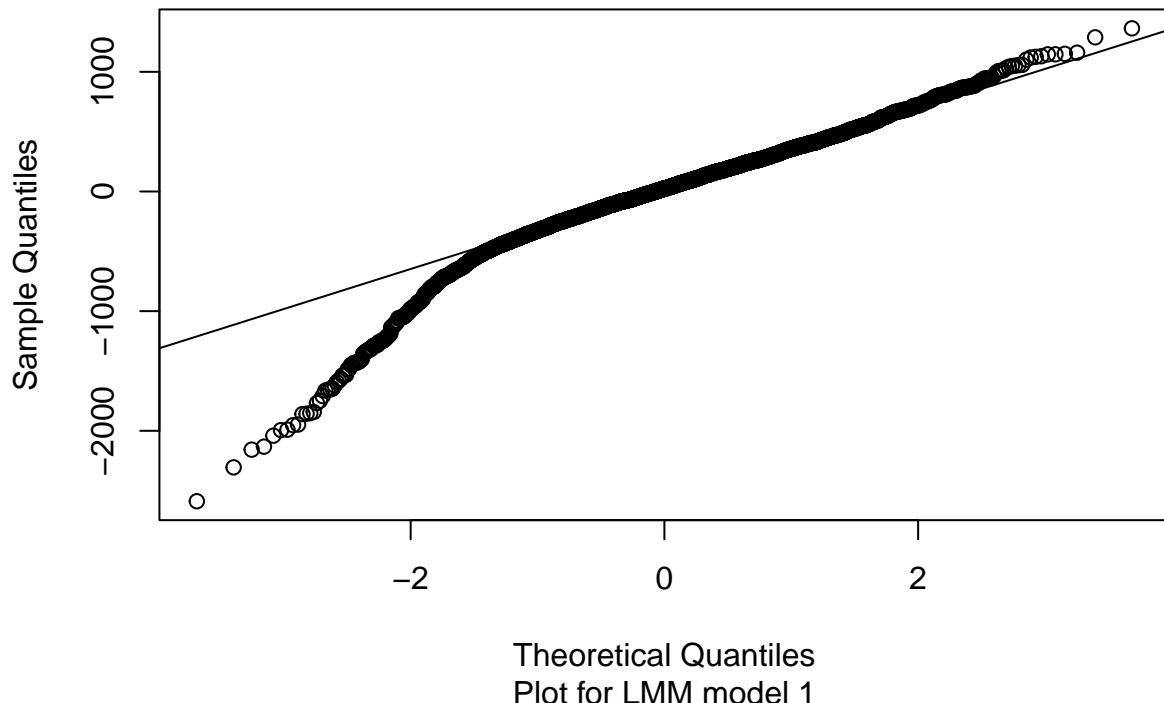
LMM 6



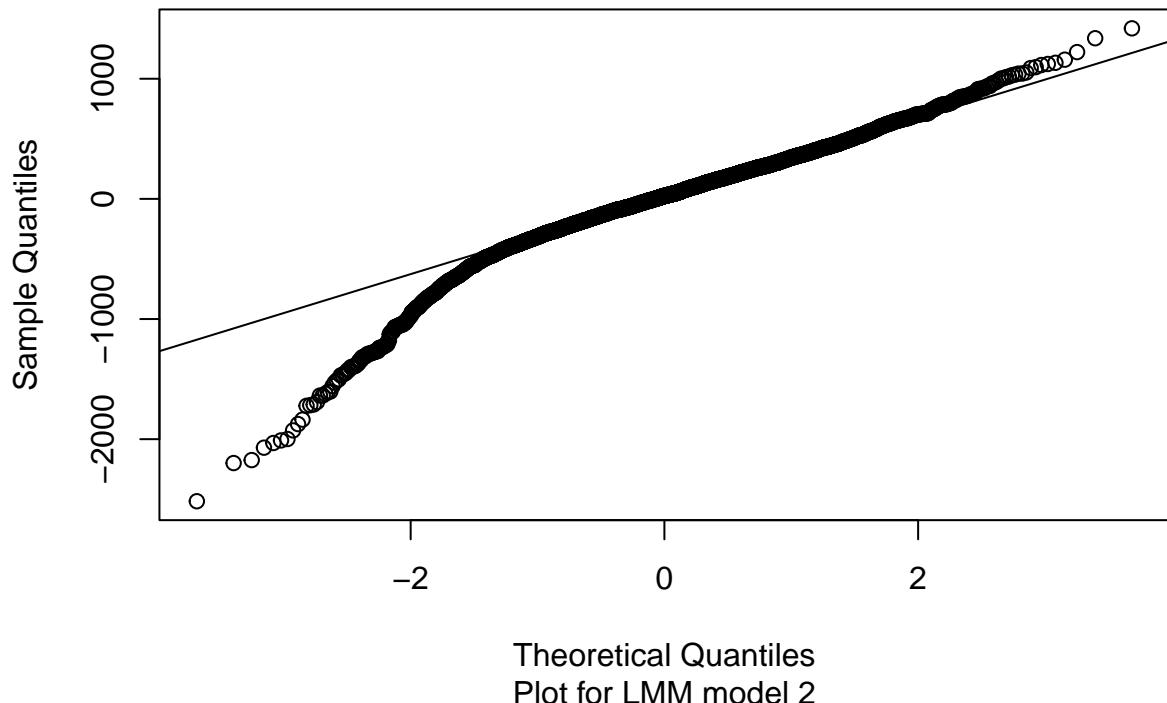
### Normal Q-Q Plot



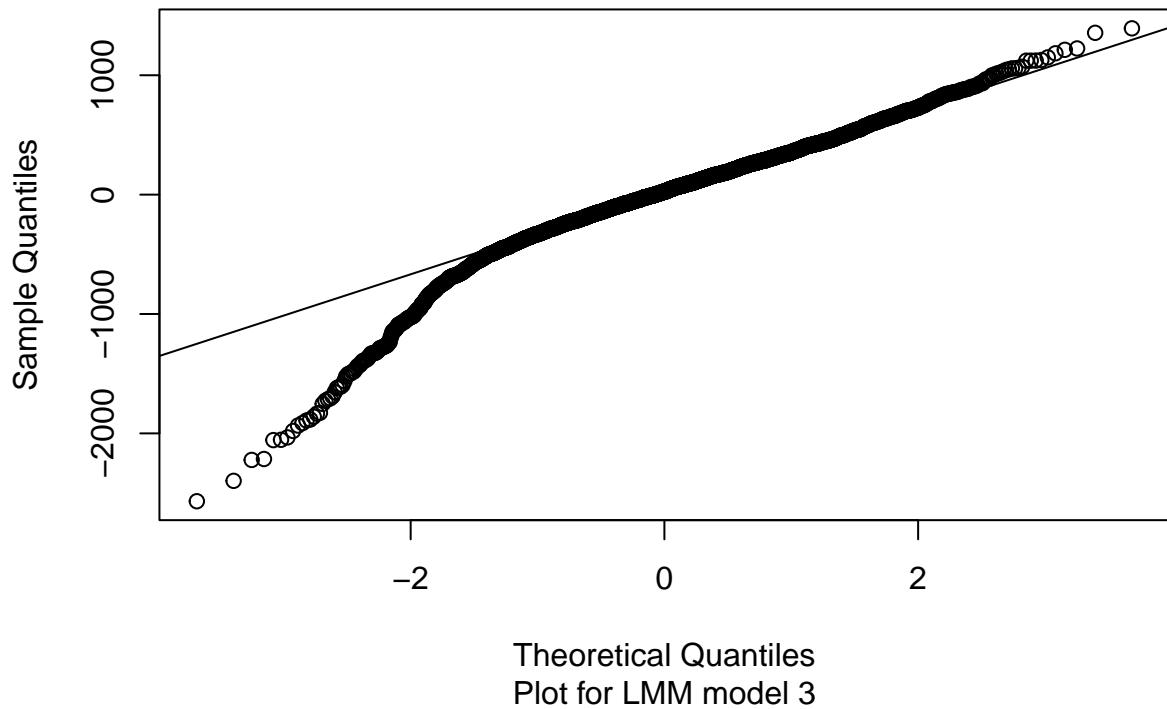
### Normal Q-Q Plot



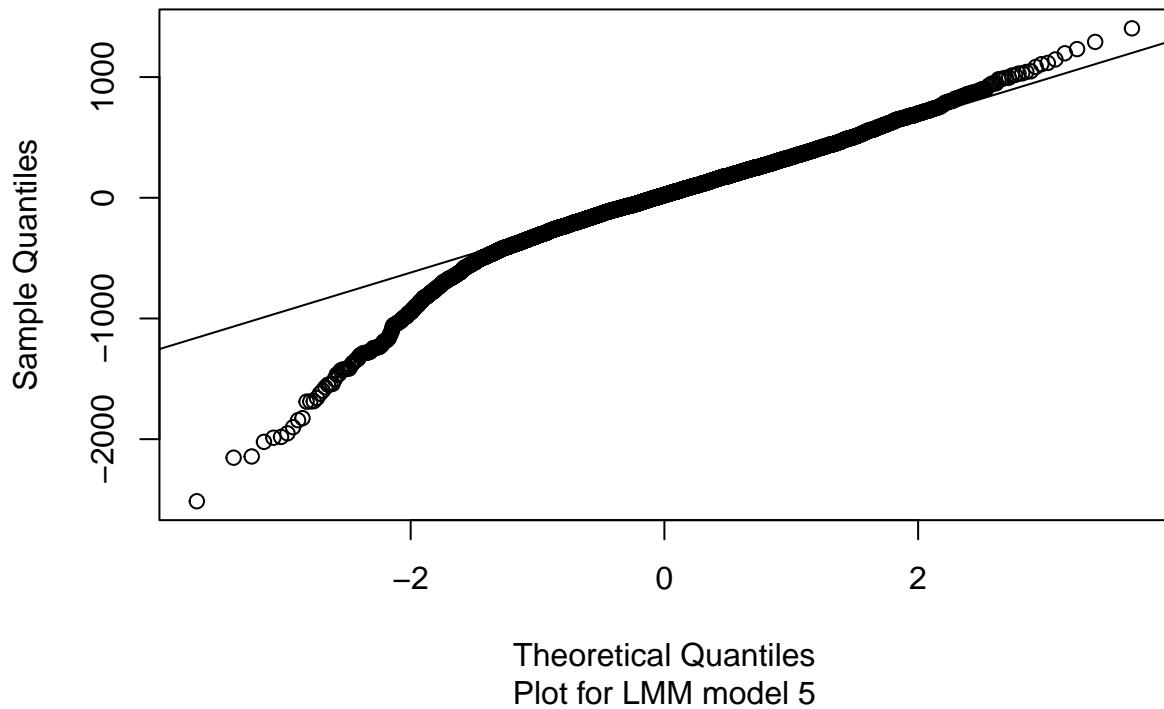
### Normal Q-Q Plot



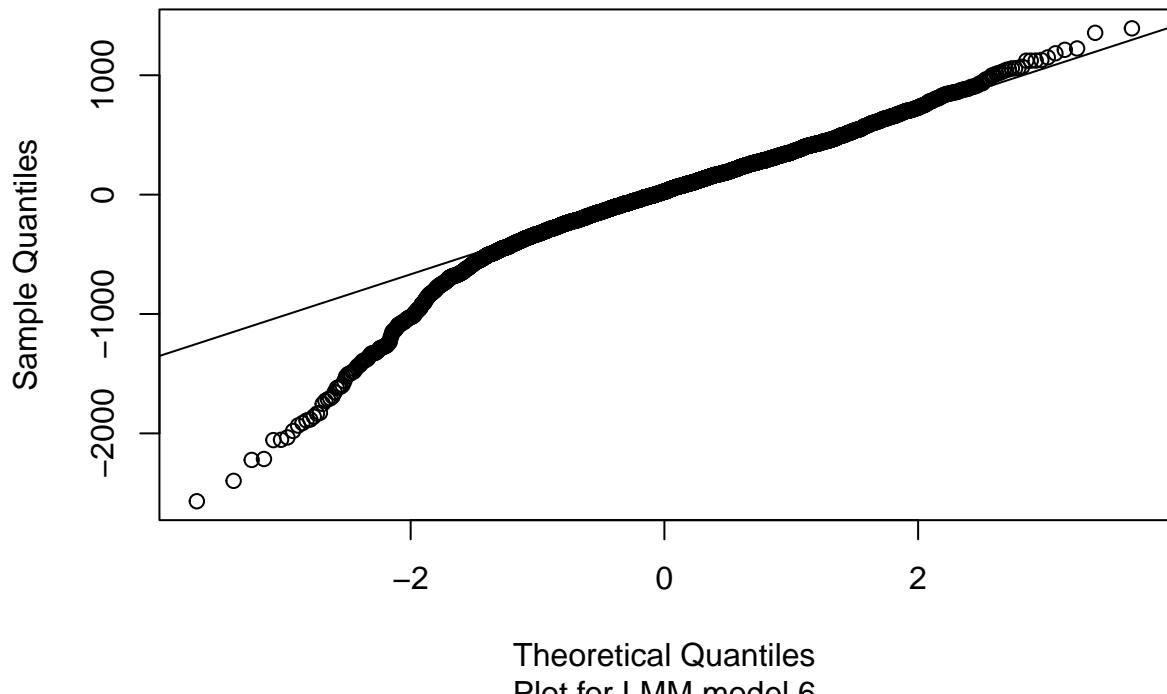
### Normal Q-Q Plot



### Normal Q-Q Plot

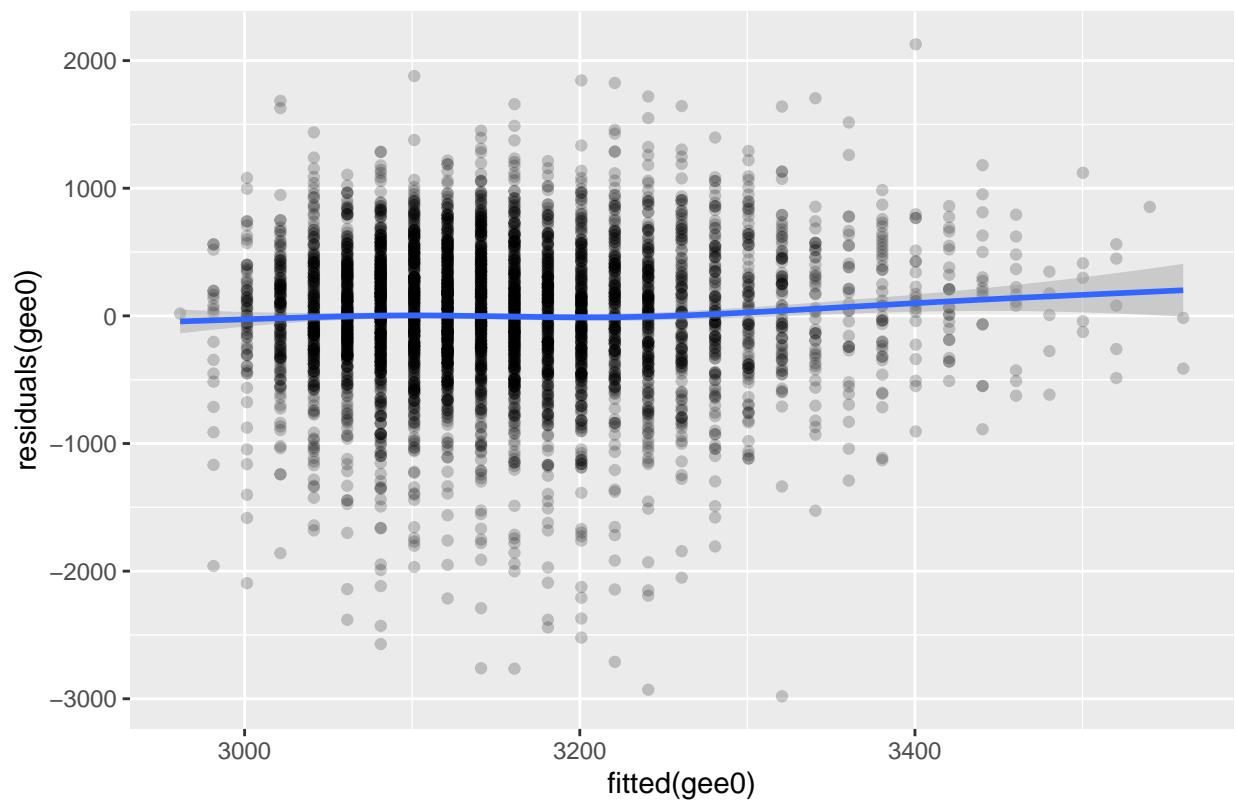


## Normal Q-Q Plot



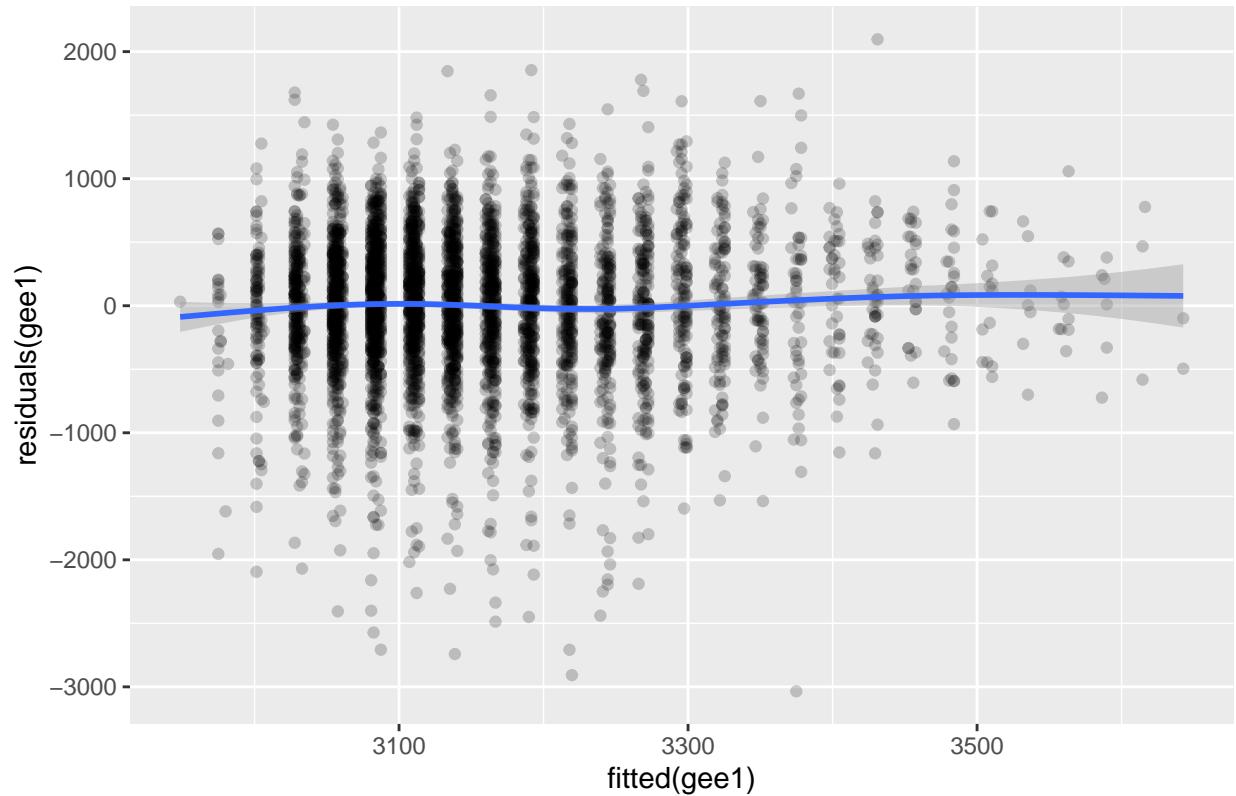
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## GEE 0



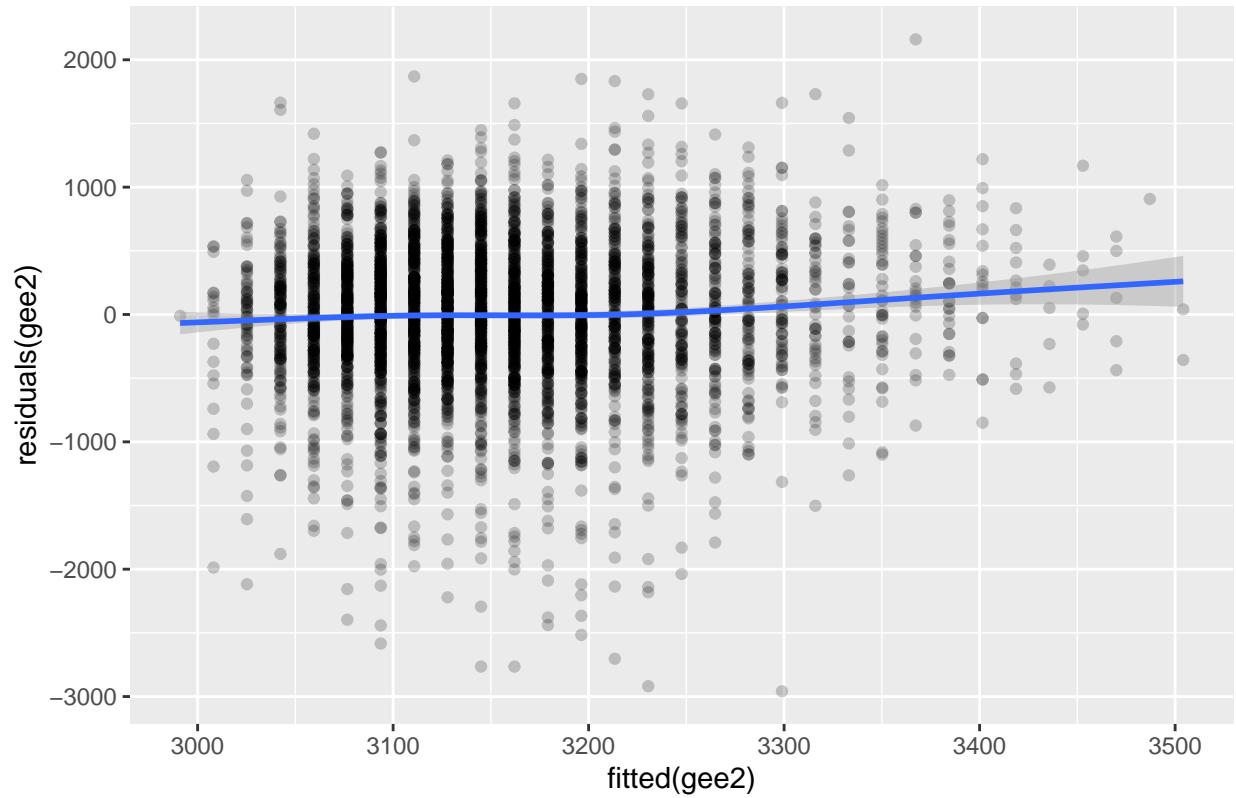
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## GEE 1



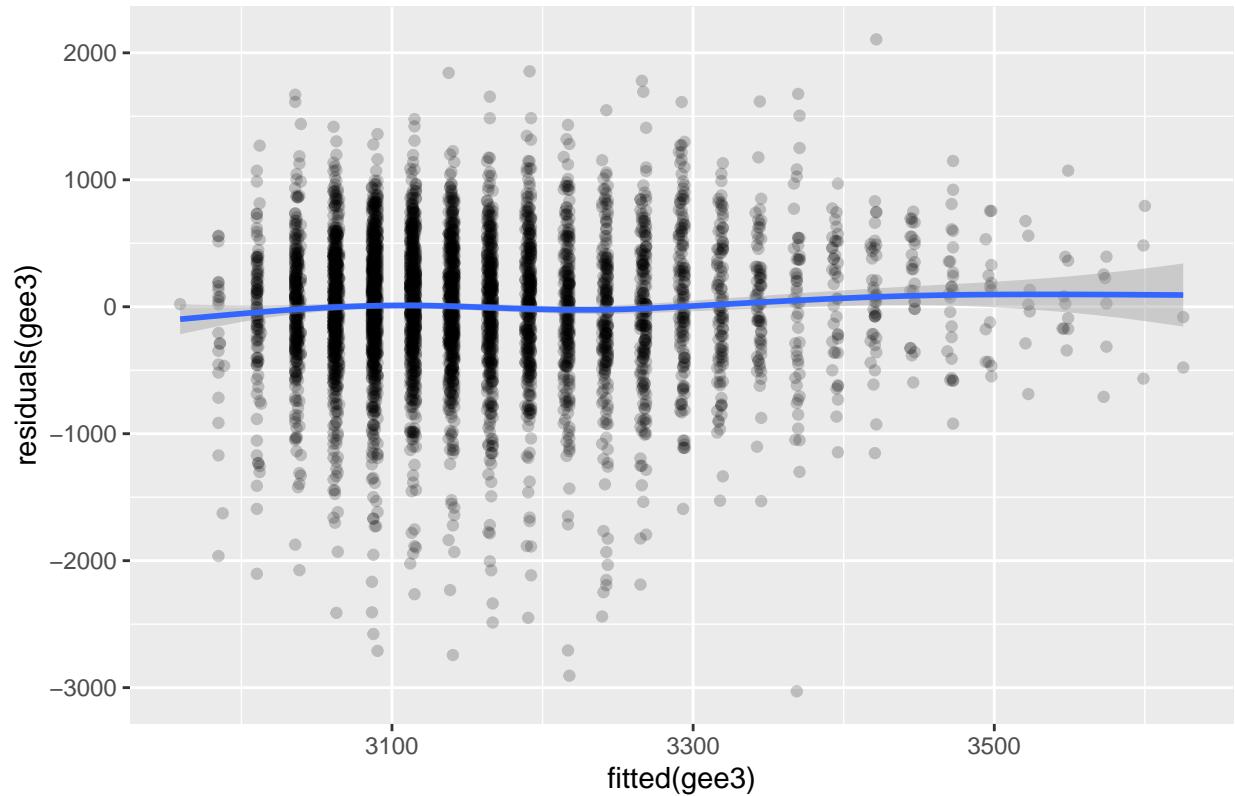
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

## GEE 2



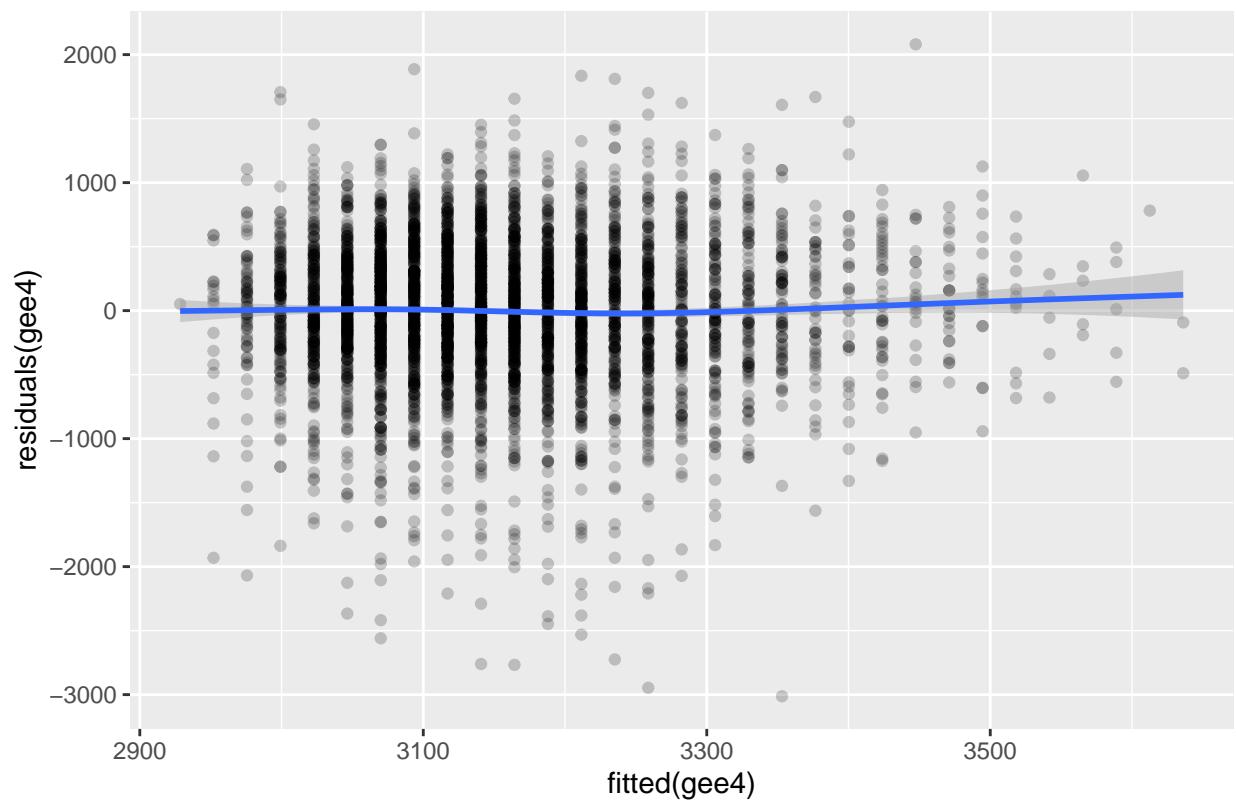
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

### GEE 3



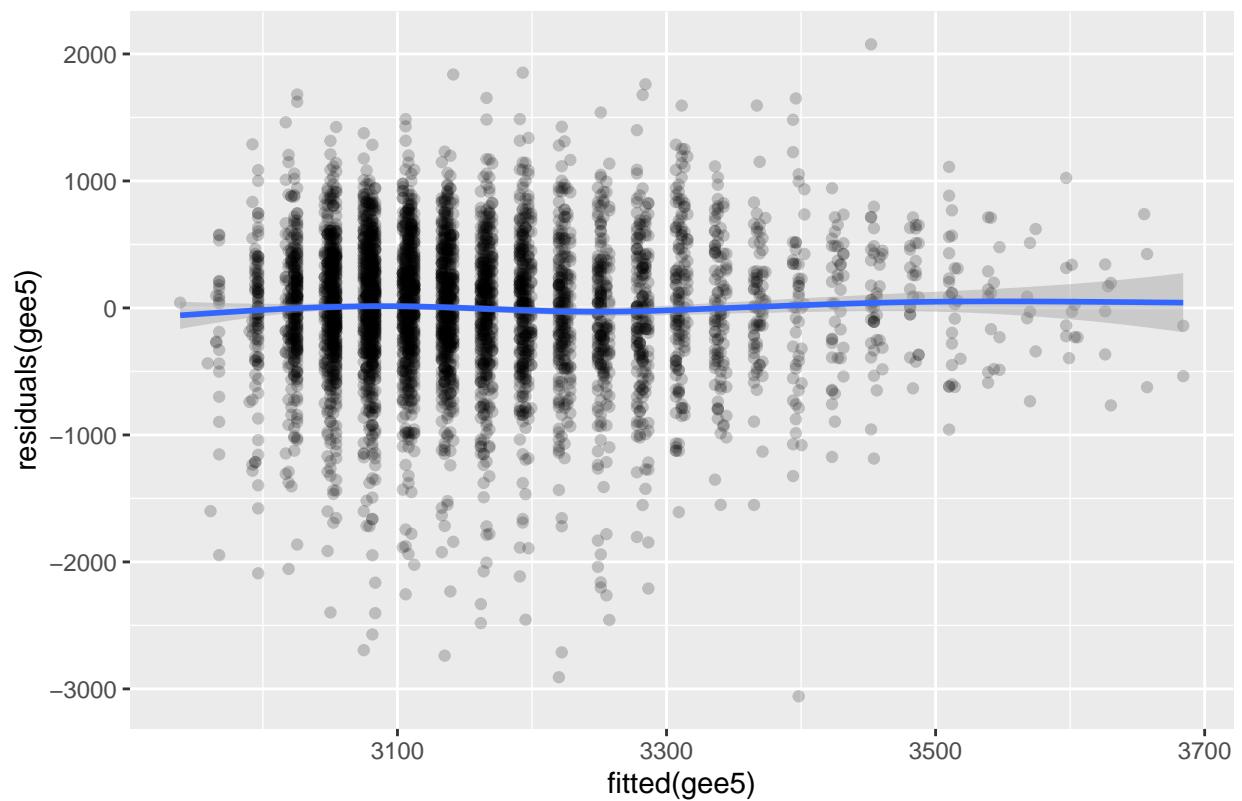
```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

GEE 4

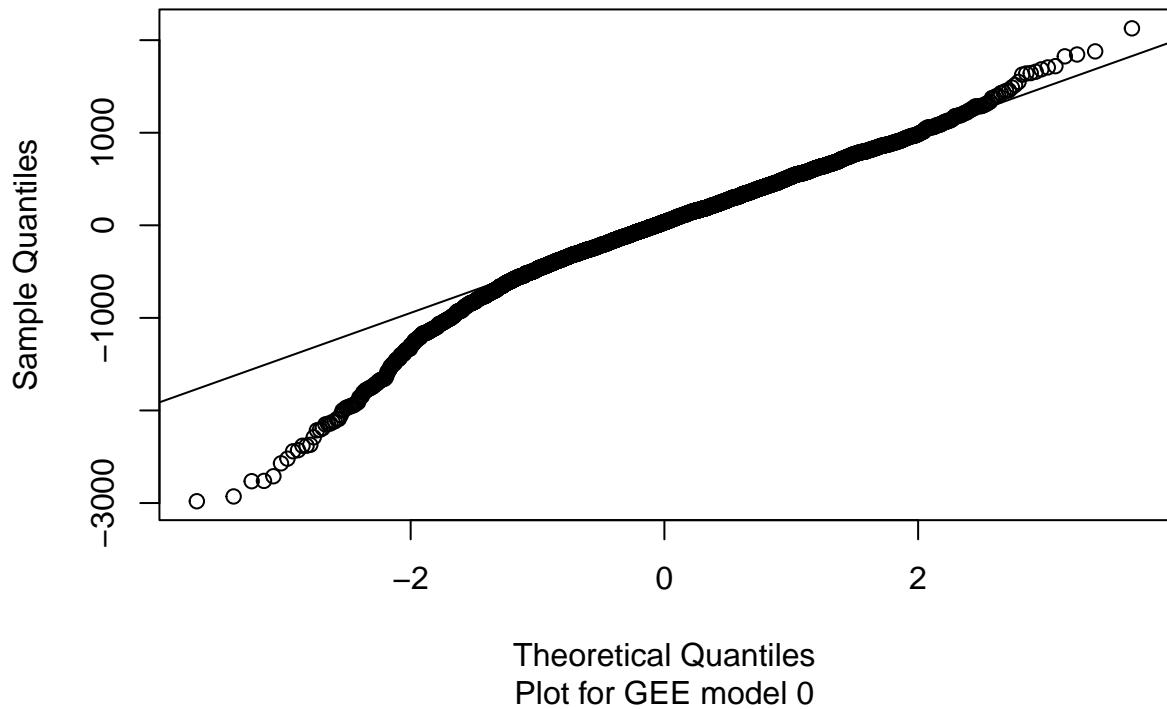


```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

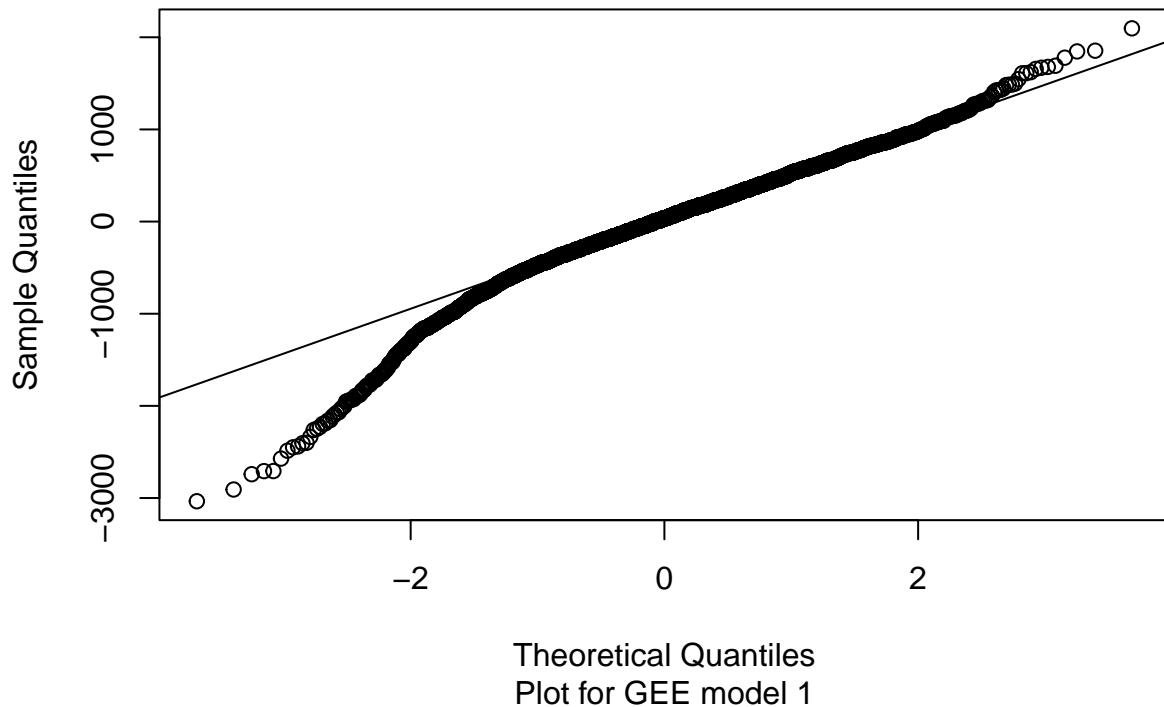
GEE 5



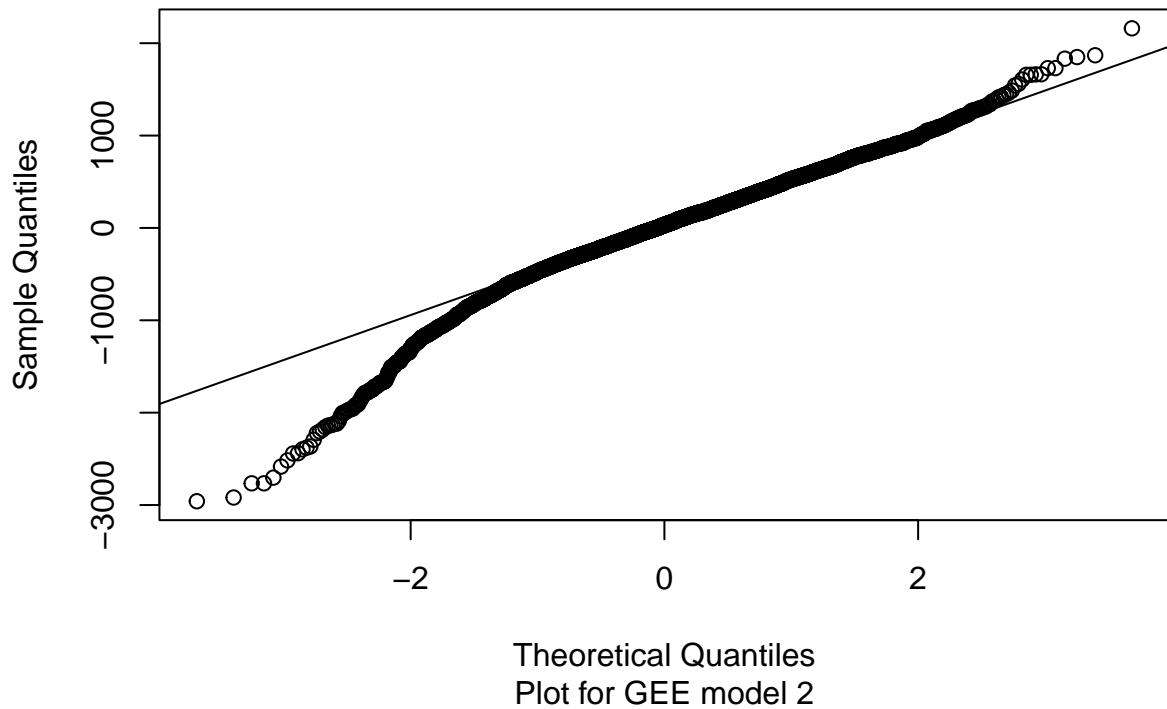
### Normal Q-Q Plot



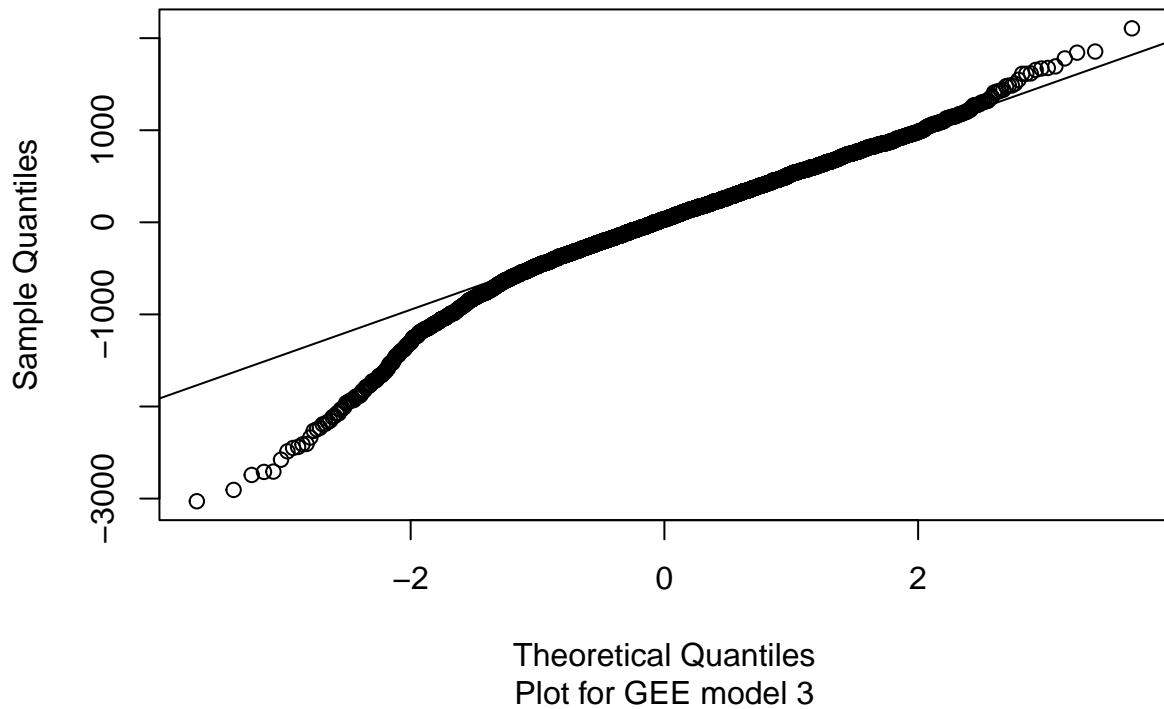
### Normal Q-Q Plot



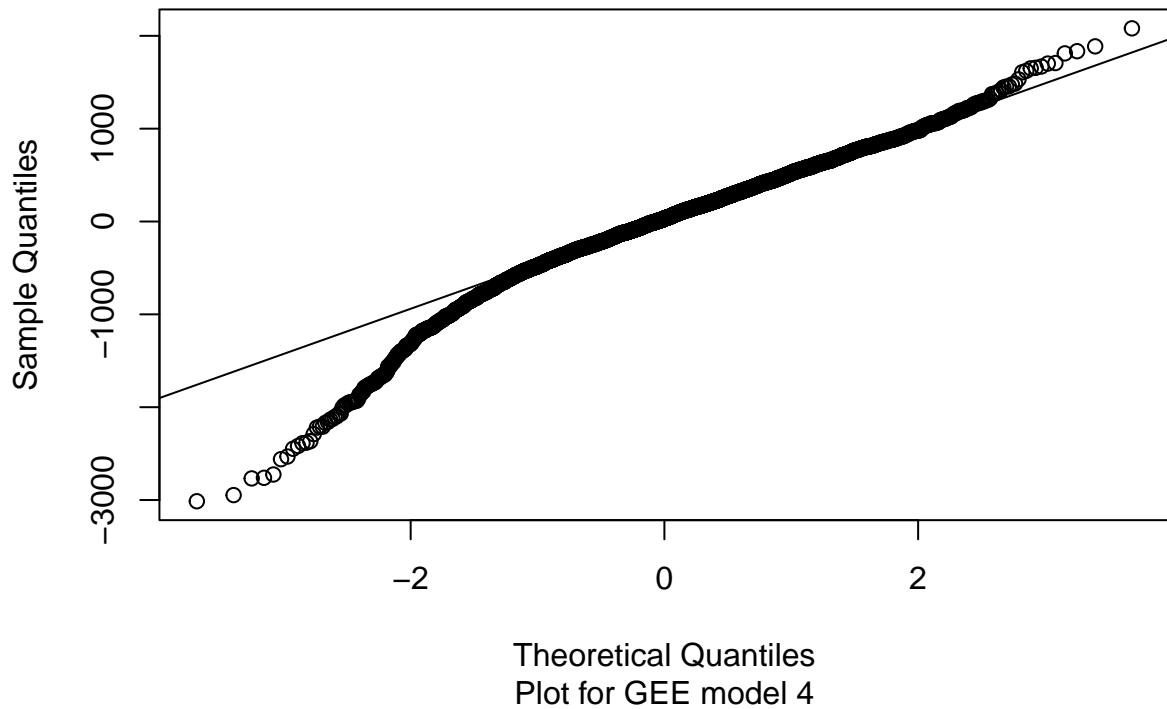
### Normal Q-Q Plot

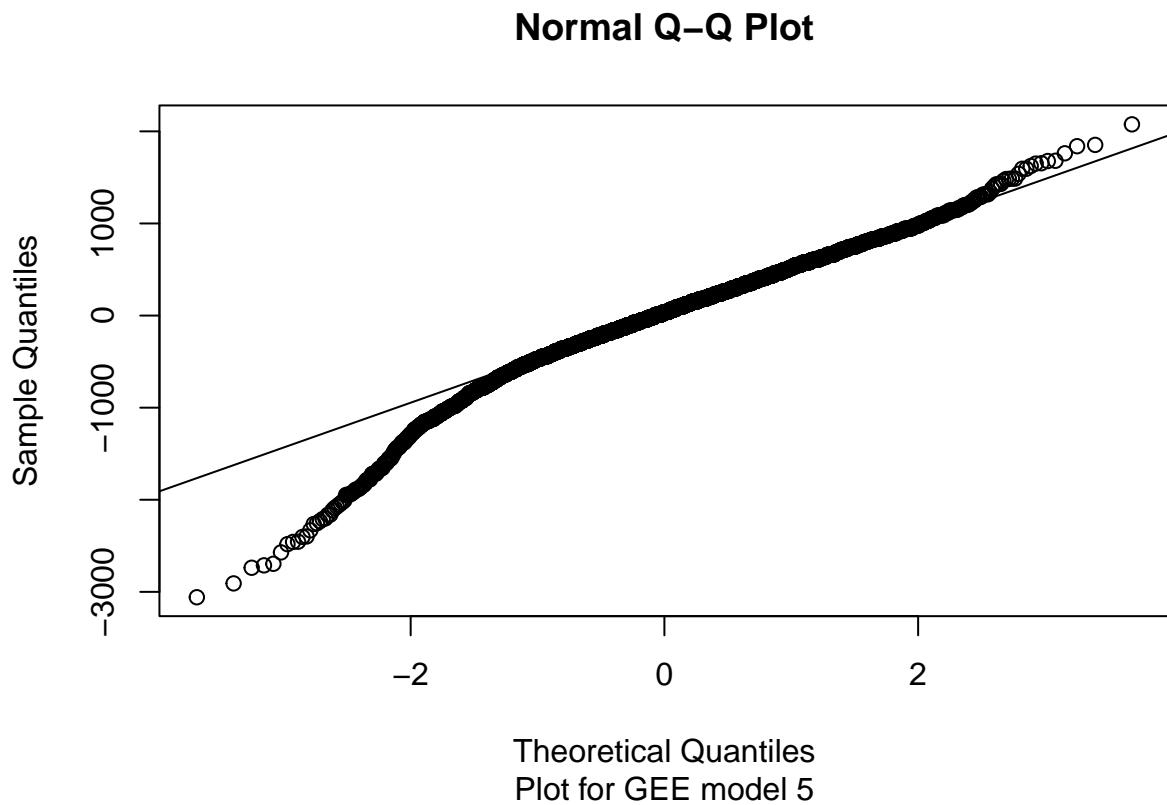


### Normal Q-Q Plot



### Normal Q-Q Plot





## Results

Table 1: Linear mixed-effects model results

Model	Effect	Estimate	CI.lower	CI.upper
Mixed	Maternal age	17.098	15.117	19.080
Mixed + Birth order + Autoreg	Maternal age	25.908	22.619	29.196
	Birth order	-24.986	-32.914	-17.057

All the LMM models agreed with one another, except for Model 0 (first row of table above) which estimated a weaker effect from maternal age than the others.

Recall,

```
# Random intercept, fixed slope; No covariates
lmm0 <- nlme::lme(birth.weight ~ maternal.age,
                    random = ~ 1 | mother.id,
                    data = birthwt)

# Random intercept and fixed slope; with birth order covariate and auto-regressive
# correlation structure
lmm3 <- lme(birth.weight ~ maternal.age + birth.order,
              random = ~ 1 | mother.id, correlation = corAR1(),
              data = birthwt)
```

Table 2: Generalized estimating equations results

Model	Effect	Estimate	CI.lower	CI.upper
Autoregressive	Maternal age	19.952	17.653	22.252
Autoregressive + Birth order	Maternal age	26.440	22.901	29.979
	Birth order	-24.757	-33.149	-16.364

Table 3: Generalized estimating equations results

Model	Effect	Estimate	CI.lower	CI.upper
Autoregressive + Birth order	Maternal age	26.440	22.901	29.979
	Birth order	-24.757	-33.149	-16.364

All the GEE models agree that adding a covariate increases the effects of age (by about 7 units). However, the other GEE models estimate different effects (I think they'll all higher). The ACF plot in the beginning convinced me that assuming auto-regressive and temporal decay of correlation is something we should stick to.

## Modeling log odds of low birth weight

NOTE I AM ASSUMING THIS IS THE CORRECT WAY TO INTERPRET GLMM'S AS SIMILAR TO LOGISTIC REGRESSION; I REALIZE NOW I MAY HAVE BEEN WRONG PLEASE VERIFY

```
# Generalized linear mixed-effects models (GLMM) to model log odds of low birth
# weight
glmm0 <- lme4::glmer(
  birth.weight.binary ~ maternal.age + (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

glmm1 <- glmer(
  birth.weight.binary ~ maternal.age + interval + (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

glmm2 <- glmer(
  birth.weight.binary ~ maternal.age + birth.order + (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

glmm3 <- glmer(
  birth.weight.binary ~ maternal.age + interval + birth.order + (1 | mother.id),
  data = birthwt %>% na.omit, family = binomial)

# Summary for all results
glmm0_res <- glmm0 %>% summary %>% coef %>% data.frame %>%
  mutate(CI.lower = Estimate + Std..Error, CI.upper = Estimate + Std..Error)

glmm1_res <- glmm1 %>% summary %>% coef %>% data.frame %>%
  mutate(CI.lower = Estimate + Std..Error, CI.upper = Estimate + Std..Error)

glmm2_res <- glmm2 %>% summary %>% coef %>% data.frame %>%
  mutate(CI.lower = Estimate + Std..Error, CI.upper = Estimate + Std..Error)

glmm3_res <- glmm3 %>% summary %>% coef %>% data.frame %>%
  mutate(CI.lower = Estimate + Std..Error, CI.upper = Estimate + Std..Error)

# Table for all results
gee_res <- rbind(glmm0_res, glmm1_res, glmm2_res, glmm3_res)[ -c(1,3,6,9), ] %>%
  exp %>% select("exp(Estimate)" = Estimate, CI.lower, CI.upper)
gee_res$Model <- c("A",
                   "B", "",
                   "C", "",
                   "D", "", "")
gee_res$Effect <- c(
  "Maternal age",
  "Maternal age", "Interval",
  "Maternal age", "Birth order",
  "Maternal age", "Interval", "Birth order")

gee_res %>%
  select(Model, Effect, everything()) %>%
  knitr::kable(row.names = FALSE, digits = 3,
               caption = "Generalized linear mixed-effects results")
```

Table 4: Generalized linear mixed-effects results

Model	Effect	exp(Estimate)	CI.lower	CI.upper
A	Maternal age	0.959	0.977	0.977
B	Maternal age	0.969	0.987	0.987
	Interval	0.848	0.908	0.908
C	Maternal age	0.939	0.961	0.961
	Birth order	1.113	1.192	1.192
D	Maternal age	0.954	0.978	0.978
	Interval	0.862	0.924	0.924
	Birth order	1.073	1.152	1.152

Exp(Estimates) is on the odds scale, please verify that if an effect's 95% CI covers 1, the we cannot conclude it has significant effect.

I don't know how to compare models through diagnostics yet.

**End of document.**