Residual Model

Jay, Alec, Lynn

5/21/2021

```
library(tidyverse)
library(nlme)
library(mgcv)
aids = read.csv("aids.csv")


aids = aids %>%
  mutate(occasion = ceiling(week),
         gender = factor(gender, level = c("male", "female")),
         treatment = as.factor(treatment))

glimpse(aids)
```

```
## Rows: 5,036
## Columns: 7
## $ id        <int> 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 4, 4, 4, 4, 5, 5, ...
## $ treatment <fct> 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 1, 3, 3, 3, 3, 1, 1, ...
## $ age       <dbl> 36.4271, 36.4271, 36.4271, 36.4271, 36.4271, 36.4271, 47....
## $ gender    <fct> male, male, male, male, male, male, male, male, male, mal...
## $ week      <dbl> 0.0000, 7.5714, 15.5714, 23.5714, 32.5714, 40.0000, 0.000...
## $ log_cd4   <dbl> 3.135494, 3.044522, 2.772589, 2.833213, 3.218876, 3.04452...
## $ occasion  <dbl> 0, 8, 16, 24, 33, 40, 0, 8, 16, 23, 31, 39, 0, 0, 8, 17, ...
```

Current Model

```
aids <- aids %>%
    mutate(knot_term1 = if_else((treatment == 4 & occasion <= 16), occasion^2, 0)) %>%
    relocate(knot_term1, .after = occasion)


ctrl <- lmeControl(opt = 'optim')
model_spline2 = lme(log_cd4 ~ occasion +
              treatment:occasion + age + knot_term1,
              data = aids,
              random = ~ occasion | id,
              method = "REML",
              control = ctrl)


summary(model_spline2)
```

```
## Linear mixed-effects model fit by REML
```

```
##   Data: aids
##        AIC      BIC    logLik
##   12086.38 12158.14 -6032.192
##
## Random effects:
##  Formula: ~occasion | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.80105770 (Intr)
## occasion    0.01572499 0.187
## Residual    0.57838872
##
## Fixed effects:  log_cd4 ~ occasion + treatment:occasion + age + knot_term1
##                           Value  Std.Error   DF   t-value p-value
## (Intercept)           2.5910590 0.11844924 3722 21.874846  0.0000
## occasion             -0.0166243 0.00169301 3722 -9.819372  0.0000
## age                   0.0103633 0.00306361 1307  3.382714  0.0007
## knot_term1            0.0014324 0.00022957 3722  6.239584  0.0000
## occasion:treatment2   0.0024427 0.00237231 3722  1.029676  0.3032
## occasion:treatment3   0.0066011 0.00236783 3722  2.787822  0.0053
## occasion:treatment4   0.0163926 0.00234301 3722  6.996385  0.0000
##  Correlation:
##                      (Intr) occasn age    knt_t1 occs:2 occs:3
## occasion             -0.030
## age                  -0.976  0.004
## knot_term1           -0.023  0.012  0.001
## occasion:treatment2   0.002 -0.704 -0.001  0.000
## occasion:treatment3   0.003 -0.705 -0.003  0.000  0.503
## occasion:treatment4   0.005 -0.713 -0.005  0.021  0.508  0.509
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.21391522 -0.43707165  0.03003612  0.48390366  3.64340496
##
## Number of Observations: 5036
## Number of Groups: 1309
```

```
model_quadratic = lme(log_cd4 ~ occasion + I(occasion^2) +
             treatment:occasion + treatment:I(occasion^2)
               +  age, data = aids,
             random = ~ occasion + I(occasion^2)| id,
             method = "REML")

summary(model_quadratic)
```

```
## Linear mixed-effects model fit by REML
##   Data: aids
##        AIC      BIC    logLik
##   12012.15 12123.03 -5989.074
##
## Random effects:
##  Formula: ~occasion + I(occasion^2) | id
##  Structure: General positive-definite, Log-Cholesky parametrization
##                StdDev       Corr
## (Intercept)   0.7599938521 (Intr) occasn
```

```
## occasion      0.0419502978  0.324
## I(occasion^2) 0.0009560662 -0.388 -0.921
## Residual      0.5471601520
##
## Fixed effects:  log_cd4 ~ occasion + I(occasion^2) + treatment:occasion + treatment:I(occasion^2) +
##                               Value  Std.Error   DF   t-value p-value
## (Intercept)                2.5752869 0.11577884 3719 22.243157  0.0000
## occasion                  -0.0121774 0.00490731 3719 -2.481477  0.0131
## I(occasion^2)             -0.0001246 0.00013405 3719 -0.929607  0.3526
## age                        0.0096541 0.00299309 1307  3.225454  0.0013
## occasion:treatment2        0.0073306 0.00687138 3719  1.066826  0.2861
## occasion:treatment3        0.0186714 0.00686433 3719  2.720053  0.0066
## occasion:treatment4        0.0438518 0.00679921 3719  6.449549  0.0000
## I(occasion^2):treatment2 -0.0001429 0.00018701 3719 -0.764175  0.4448
## I(occasion^2):treatment3 -0.0003638 0.00018841 3719 -1.931005  0.0536
## I(occasion^2):treatment4 -0.0008388 0.00018542 3719 -4.523868  0.0000
##  Correlation:
##                           (Intr) occasn I(c^2) age    occs:2 occs:3 occs:4
## occasion                  -0.026
## I(occasion^2)              0.016 -0.939
## age                       -0.975  0.003 -0.002
## occasion:treatment2        0.002 -0.707  0.666 -0.001
## occasion:treatment3        0.002 -0.707  0.667 -0.002  0.505
## occasion:treatment4        0.002 -0.714  0.673 -0.001  0.510  0.510
## I(occasion^2):treatment2 -0.002  0.669 -0.714  0.001 -0.939 -0.478 -0.482
## I(occasion^2):treatment3 -0.002  0.663 -0.709  0.002 -0.474 -0.939 -0.479
## I(occasion^2):treatment4  0.000  0.674 -0.720 -0.001 -0.481 -0.482 -0.939
##                           I(^2):2 I(^2):3
## occasion
## I(occasion^2)
## age
## occasion:treatment2
## occasion:treatment3
## occasion:treatment4
## I(occasion^2):treatment2
## I(occasion^2):treatment3  0.508
## I(occasion^2):treatment4  0.516   0.512
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.48485816 -0.41732683  0.02411476  0.46989002  3.88629654
##
## Number of Observations: 5036
## Number of Groups: 1309
```

```
#Anova on our two models above
anova(model_spline2, model_quadratic)
```

```
## Warning in anova.lme(model_spline2, model_quadratic): fitted objects with
## different fixed effects. REML comparisons are not meaningful.
```

```
##                 Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model_spline2       1 11 12086.38 12158.14 -6032.192
## model_quadratic     2 17 12012.15 12123.03 -5989.074 1 vs 2 86.23638  <.0001
```

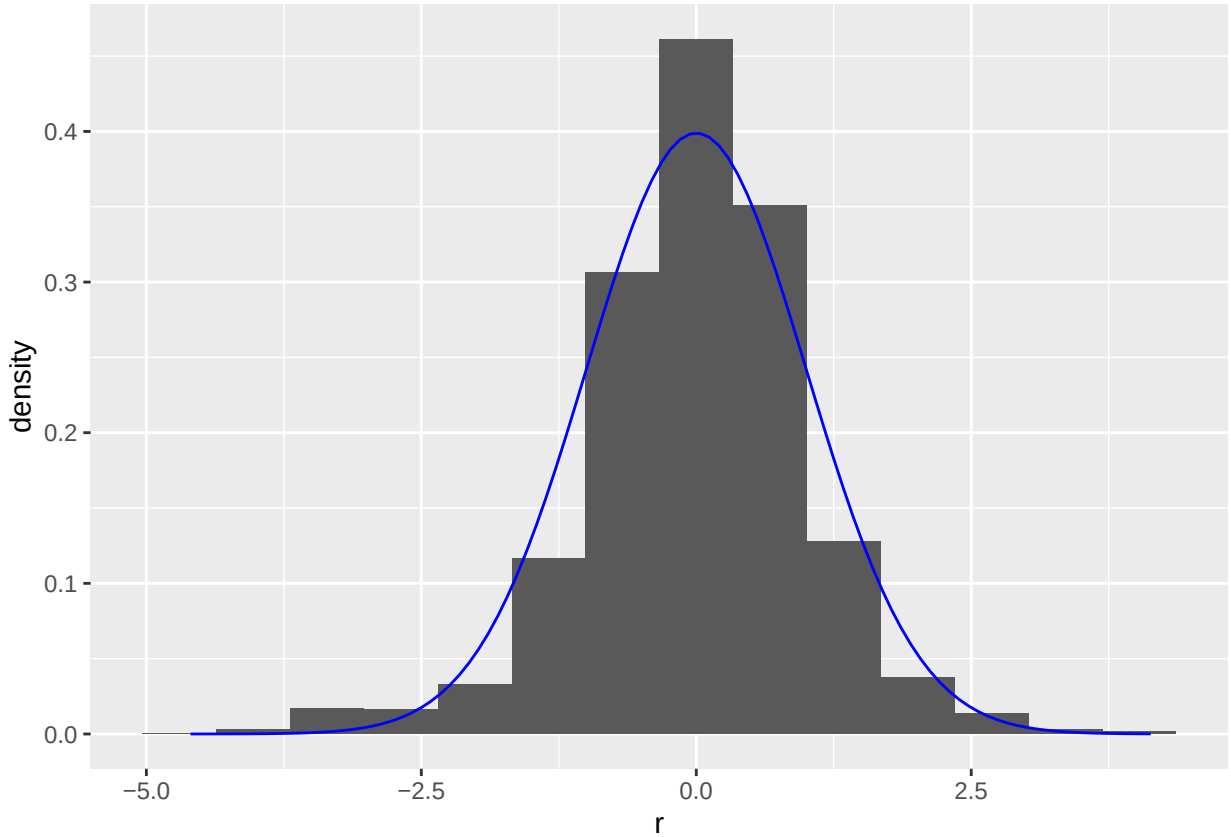
An ANOVA analysis on our two models above demonstrates that the quadratic model seems to be a better

fit for the data that we are given. A significant p-value ($p < 0.001$) further verifies this assertion.

Residual Analysis

Histogram of Transformed and Non-transformed Residuals

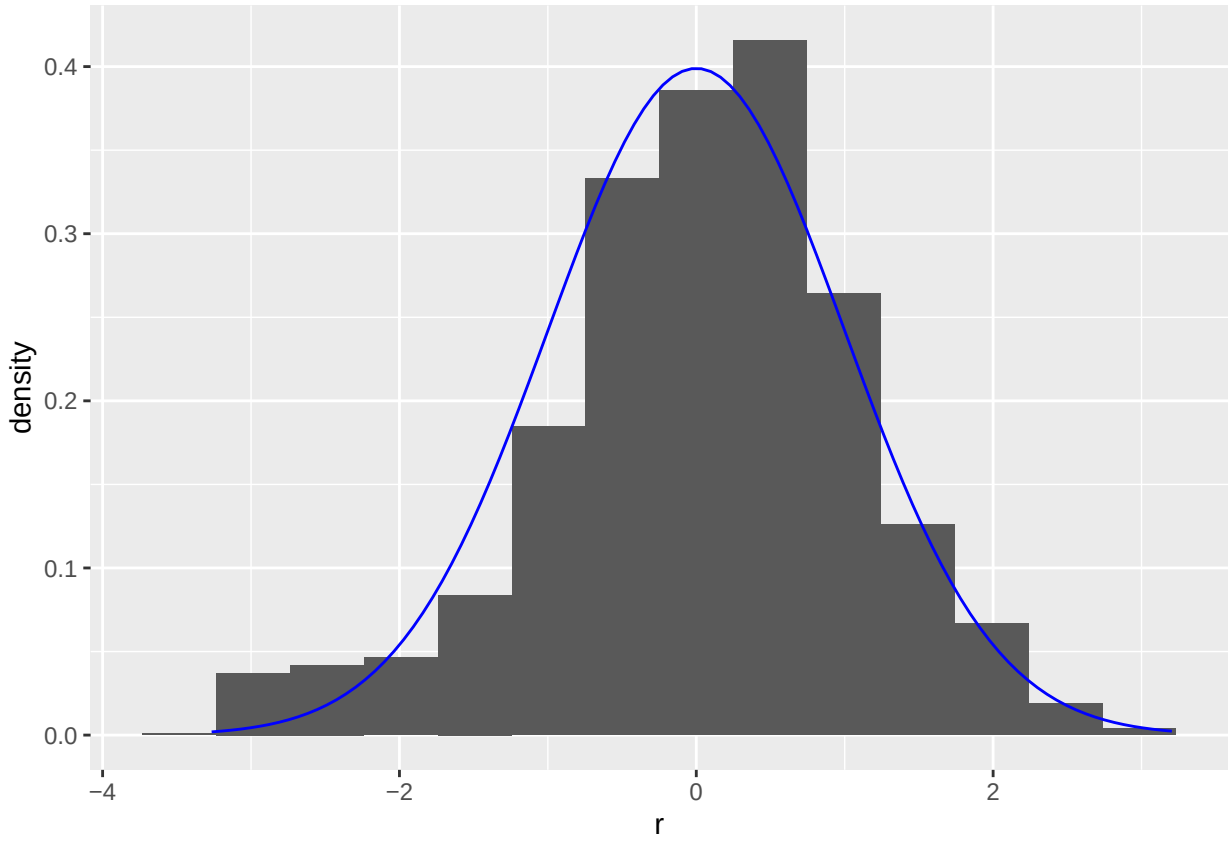
```
res_prog = residuals(model_quadratic, level = 0)


sigma_i = extract.lme.cov(model_quadratic, aids)

#lower triangular matrix
L_i = t(chol(sigma_i))

#transformed residuals
res_prog_trans = solve(L_i) %*% res_prog


tibble(r = res_prog_trans) %>%
  ggplot(aes(x = r)) +
  geom_histogram(aes(y = stat(density)), bins = 14 ) +
  geom_function(fun = dnorm, color = "blue")
```



This plot shows the density of our transformed residuals. It appears that our transformed residuals seem to follow a Normal distribution, or at the very least, a a symmetric distribution.

```
tibble(r = res_prog) %>%
  ggplot(aes(x = r)) +
```

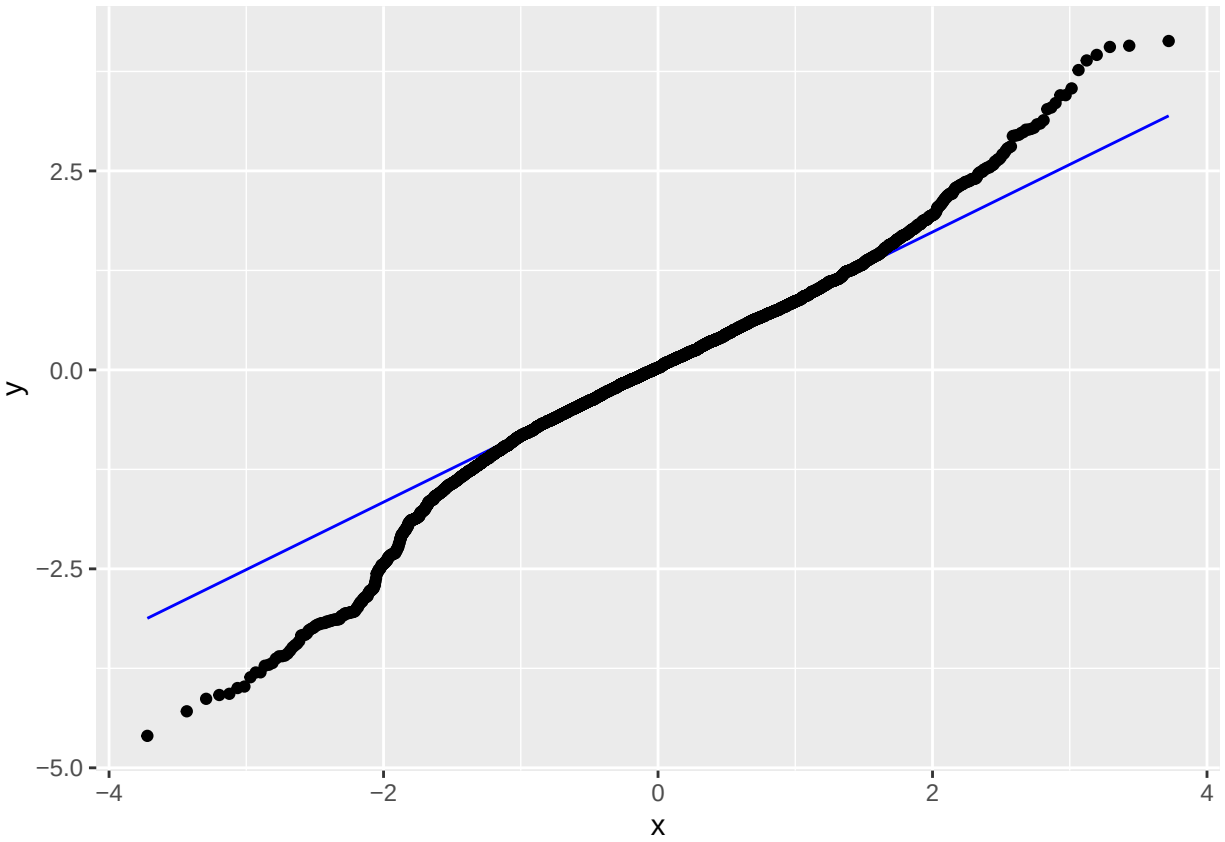
```
  geom_histogram(aes(y = stat(density)), bins = 14 ) +
  geom_function(fun = dnorm, color = "blue")
```



This plot shows the distribution of our residuals (without any transformation). The distribution appears to be a little more left-skewed in comparison to the histogram of our transformed residuals.

QQ-Plot

```
tibble(r = res_prog_trans) %>%
  ggplot(aes(sample = r)) +
  geom_qq_line(color = "blue") +
  geom_qq()
```



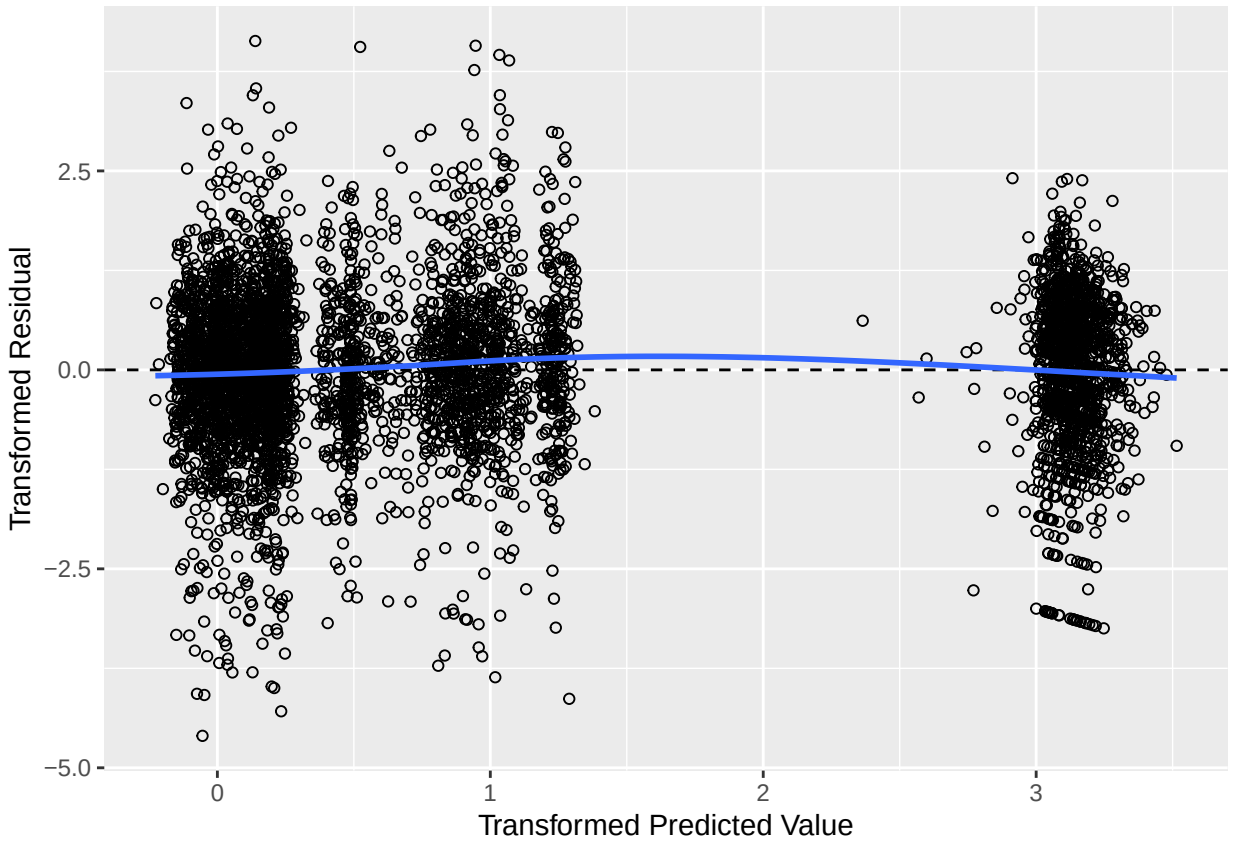
<<<<<<< HEAD The qq plot reveals the tails for the distribution of our residuals are quite heavy. This brings into question whether or not the residuals actually follow a normal distribution. Furthermore, several lingering points at the end of each tail suggest that we may have outliers present in our dataset (this is further discussed in the mahalanobis data section below).

=======

Transformed Predicted values vs. Transformed Residuals

```
#transformed vs actual residual
mu_hat = fitted(model_quadratic, level = 0)
mu_hat_transformed = solve(L_i) %*% mu_hat


tibble(x = mu_hat_transformed, y = res_prog_trans) %>%
  ggplot(aes(x = x, y = y)) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  geom_point(shape = 1) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Transformed Predicted Value", y = "Transformed Residual")
```

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

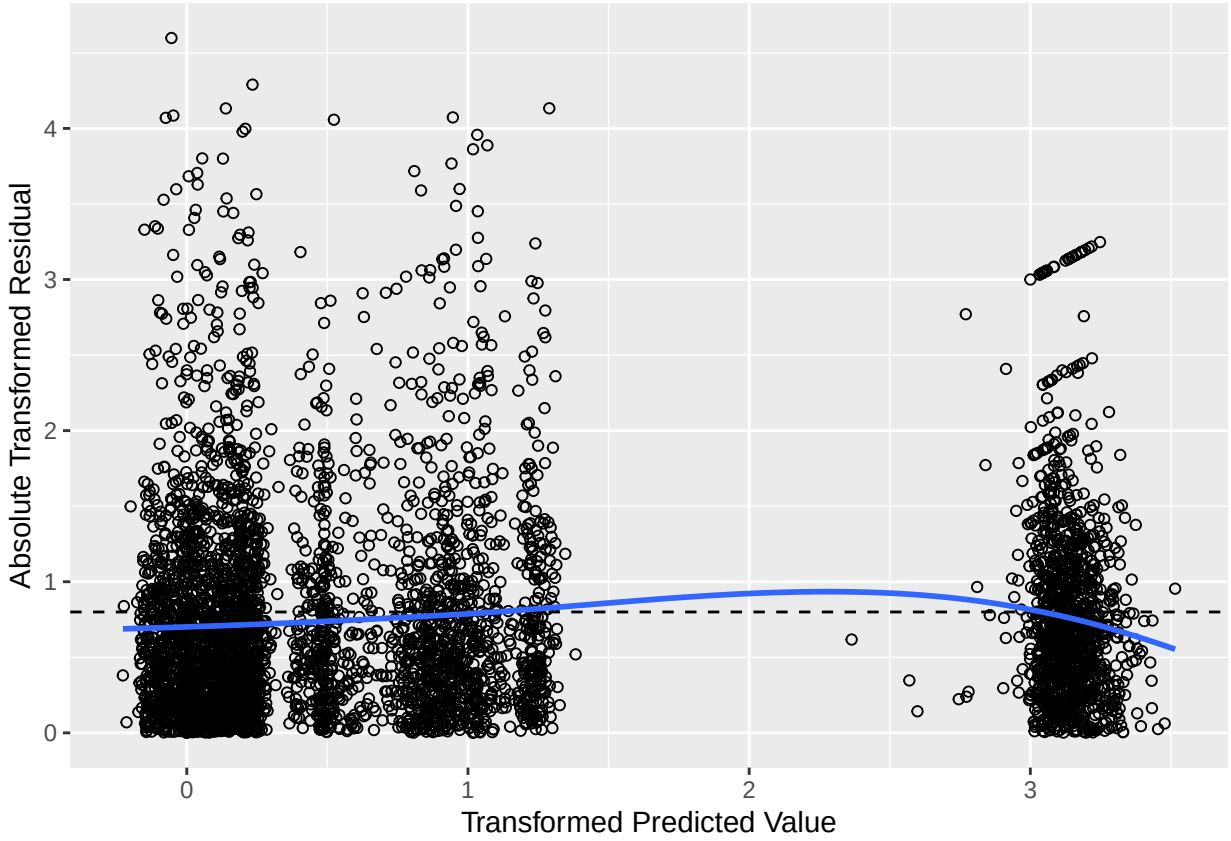


We can see that there doesn't seem to be any significant curvature in this graph, indicating that the constant variance assumption is correct.
However, we can see that there is a large gap between the points, which shows that there seems to be a gap in observed the covariates (x values).

Transformed Predicted values vs. Absolute Transformed Residuals

```
#pred_prog_trans_abs = abs(mu_hat_transformed)
res_prog_trans_abs = abs(res_prog_trans)

tibble(x = mu_hat_transformed, y = res_prog_trans_abs) %>%
  ggplot(aes(x = x, y = y)) +
  geom_hline(yintercept = 0.8, linetype = "dashed") +
  geom_point(shape = 1) +
  geom_smooth(method = "loess", se = FALSE) +
  labs(x = "Transformed Predicted Value", y = "Absolute Transformed Residual")
```

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



Using the loess smoothed curve, we can see that there is not a noticable departure from the straight line centered at approximately 0.8. The smoothed curve is relatively straight, but is centered around 0.8. This indicates that the model for the variance is adequate.

Mahalanobis Data

```
mahalanobis_data = tibble(id = aids$id, r_star = res_prog_trans) %>% group_by(id) %>%
  nest()

mahalanobis_data = mahalanobis_data %>%
  mutate(df = map_dbl(data, ~nrow(.x)))

mahalanobis_dist = function(x){
  x = as.matrix(x)
  t(x) %*% x
}

mahalanobis_data = mahalanobis_data %>%
  mutate(d = map_dbl(data, ~mahalanobis_dist(.x)))

mahalanobis_data = mahalanobis_data %>%
  mutate(p_value = pchisq(d, df, lower.tail = FALSE))


mahalanobis_data_p = mahalanobis_data %>%
  arrange(p_value)
```

```
mahalanobis_data_p %>% filter(p_value <=.05)
```

```
## # A tibble: 129 x 5
## # Groups:   id [129]
##       id data                df     d      p_value
##    <int> <list>           <dbl> <dbl>        <dbl>
##  1   178 <tibble [5 x 1]>     5  43.3 0.0000000324
##  2   692 <tibble [5 x 1]>     5  38.0 0.000000376
##  3  1118 <tibble [5 x 1]>     5  34.5 0.00000191
##  4  1193 <tibble [4 x 1]>     4  30.3 0.00000417
##  5  1100 <tibble [3 x 1]>     3  24.6 0.0000190
##  6  1207 <tibble [5 x 1]>     5  29.3 0.0000198
##  7   362 <tibble [5 x 1]>     5  28.7 0.0000271
##  8  1110 <tibble [6 x 1]>     6  30.5 0.0000315
##  9   877 <tibble [6 x 1]>     6  30.3 0.0000347
## 10   479 <tibble [6 x 1]>     6  30.1 0.0000374
## # ... with 119 more rows
```

```
#expected outliers are

expected = 5036 *.05
expected
```

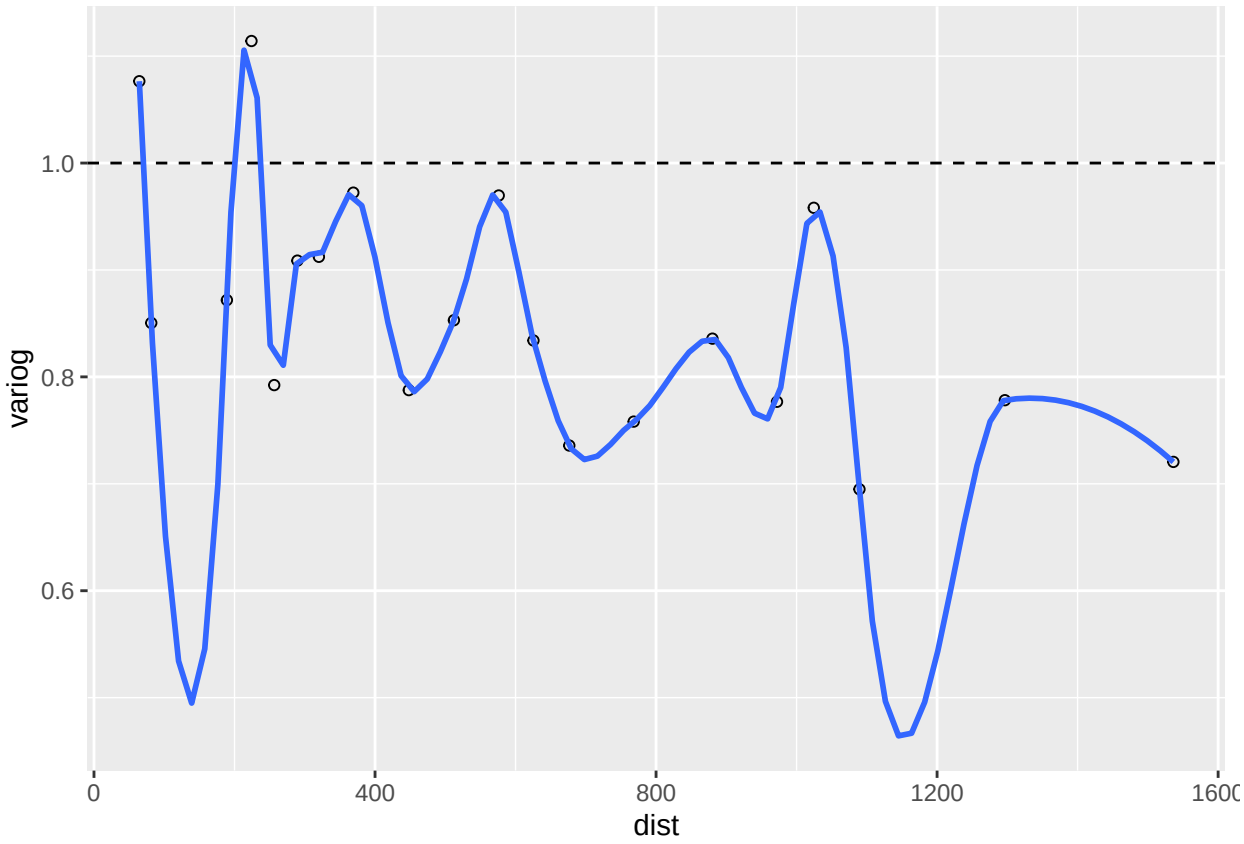
```
## [1] 251.8
```

Using the mahalanobis data, we can see that we have 129 subjects who have a p-value < 0.05. From the size of our data, we expect that we will have 251.8 outliers. Our actual 129 outliers fall within the range of 251.8, so we do not need to be concerned about the outliers we find here and see in the QQ-plot.

Semi-Variogram

```
aids = aids %>%
  mutate(occasion_sqr = occasion ^ 2)

Variogram(model_quadratic,
          data = aids,
          form = ~ occasion + occasion_sqr| id,
          resType = "normalized") %>%
  as_tibble() %>%
  ggplot(aes(x = dist, y = variog)) +
  geom_hline(yintercept = 1, linetype = "dashed") +
  geom_point(shape = 1) +
  geom_smooth(method = "loess", se = FALSE, span = .2)
```

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



Looking at the semi-variogram, we can see that the loess smoothed curve does seem to fluctuate randomly around 1.0, but has a general decreasing trend. This could indicate that our model's covariance matrix may not be adequate.