In Class Work

Trevor, Johnathan, Grace, Ava

December 03, 2021

Background

The code below reads in a dataset with public health data from Zambia and does some data cleaning. The response variable of interest is **height_zscore**, the z-score of the child's height compared to the national average. (The assumption is that malnourished or unhealthy children will be unusually small.) Other variables include:

- **child_gender**
- **breastf** duration of breast-feeding in months
- **child_age** child's age in months
- **mother_birth_age** mother's age when the child was born, in years
- **mother_height** mother's height in cm
- **mother_BMI** mother's body mass index
- **mother_education** mother's education level
- **mother_work** mother's work status
- **region** Region in Zambia of mother's residence
- **district** District in Zambia of mother's residence

```
zam <- read.table('http://www.uni-goettingen.de/de/document/download/d90a2d7b26c4504ab6630cf36cbae2fa.ra
                  header=TRUE)
names(zam) <- c('height_zscore', 'child_gender', 'breastf', 'child_age',
                'mother_birth_age', 'mother_height', 'mother_BMI',
                'mother_education', 'mother_work', 'district', 'region', 'time')
zam <- zam %>% mutate(child_gender = ifelse(child_gender==1, 'Male', 'Female')) %>%
  mutate(mother_education = factor(mother_education)) %>%
  mutate(mother_education = fct_recode(mother_education,
                                       'None' = '1',
                                       'Primary School' = '2',
                                       'Secondary School' = '3',
                                       'Higher Education' = '4')) %>%
  mutate(mother_work = ifelse(mother_work==1, 'Working', 'Not Working')) %>%
  mutate(region = factor(region)) %>%
  mutate(region = fct_recode(region,
                             'Central' = '1',
                             'Copperbelt' = '2',
                             'Eastern' = '3',
                             'Luapula' = '4',
                             'Lusaka' = '5',
                             'Northern' = '6',
                             'Northwestern' = '7',
                             'Southern' = '8',
                             'Western' = '9')) %>%
  mutate(district = factor(district)) %>%
```

```
  dplyr::select(-time)
zam <- arrange(zam, district)
glimpse(zam)
```

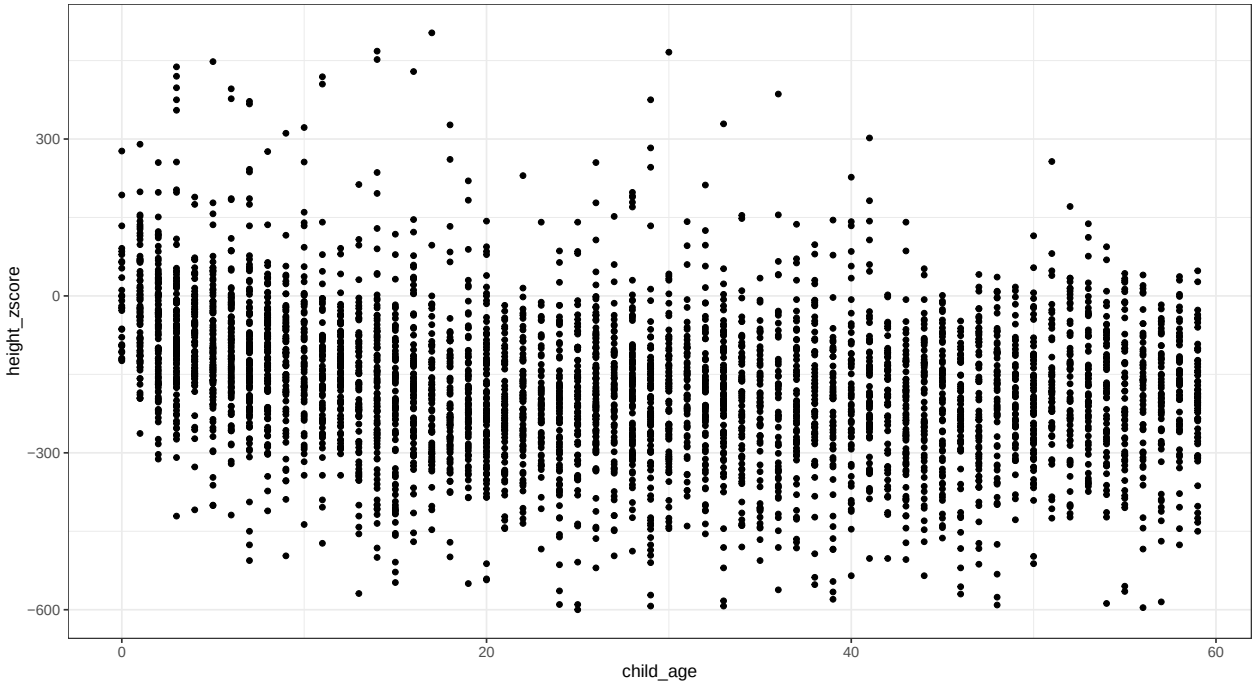
```
## Rows: 4,421
## Columns: 11
## $ height_zscore    <int> -264, -389, -127, -169, -156, -269, -169, 5, -279, 10~
## $ child_gender     <chr> "Male", "Female", "Female", "Male", "Male", "Female",~
## $ breastf          <int> 24, 19, 1, 24, 0, 0, 16, 14, 19, 11, 1, 40, 21, 21, 0~
## $ child_age        <int> 29, 57, 16, 46, 9, 5, 30, 56, 25, 13, 16, 46, 32, 33,~
## $ mother_birth_age <dbl> 25.58333, 23.25000, 35.66667, 33.16667, 31.25000, 35.~
## $ mother_height    <dbl> 162.4, 162.4, 151.8, 151.8, 156.6, 161.1, 161.1, 161.~
## $ mother_BMI       <dbl> 22.33, 22.33, 18.66, 18.66, 24.22, 25.58, 25.58, 25.5~
## $ mother_education <fct> Primary School, Primary School, Primary School, Prima~
## $ mother_work      <chr> "Working", "Working", "Working", "Working", "Working"~
## $ district         <fct> 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 12, 1~
## $ region           <fct> Northern, Northern, Northern, Northern, Northern, Nor~
```

Questions

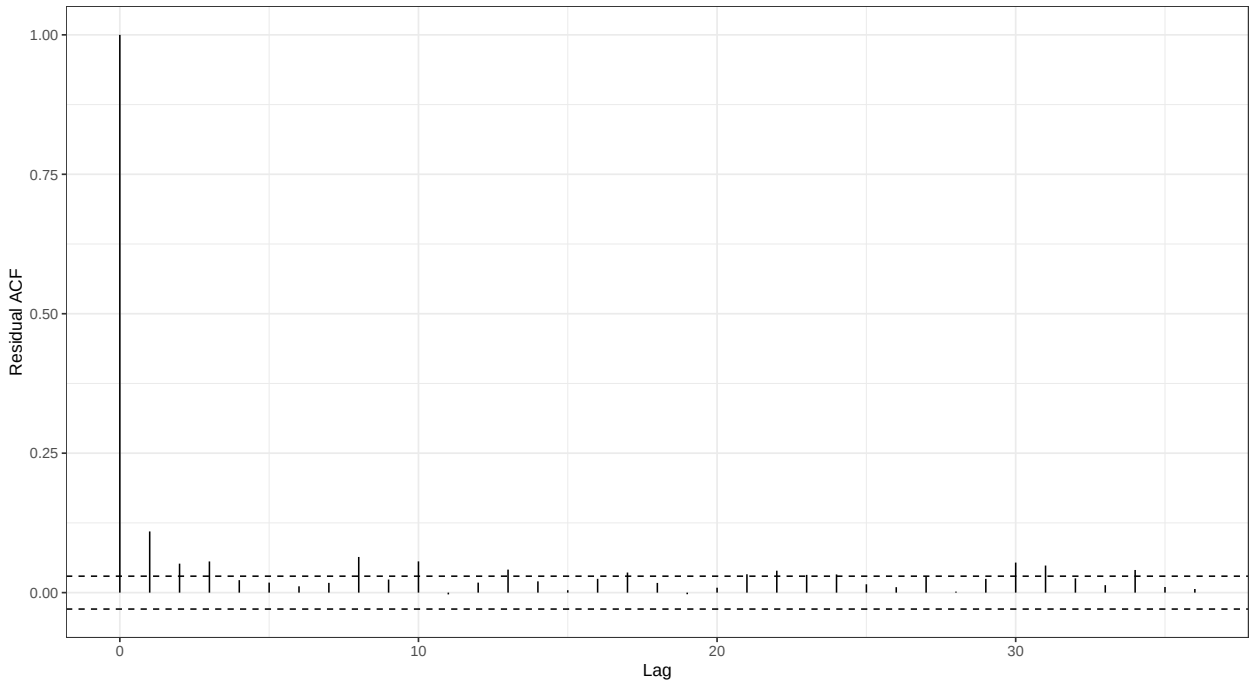
- Plan a regression model with `height_zscore` as the response variable. Discuss with your group which predictor(s) will be smooth, and what your choices of basis function and `k` will be. Are there other variables that you wish you had in the dataset so that you could include them as predictors (what)?

```
model <- glmmTMB(height_zscore ~ mother_height + mother_BMI + child_age, data = zam)
```

```
gf_point(height_zscore ~ child_age, data = zam)
```




```
gf_acf(~model)
```



- Do exploratory data analysis to familiarize yourself further with the data
- Fit a GAM to this dataset (with `height_zscore` as the response variable). View the summary and maybe the `gam.check()` to make sure everything looks OK (no warnings, failure to converge, NAs, etc.).
- *Skip this section for now unless you have at least 10-15 minutes left. You can return at the end if time permits.* What conditions do you need to check for your model? Make model assessment plots and check them.
- Make prediction plots for the expected `height_zscore` as a function of two or more of your predictors (prioritizing smooth terms). What patterns do you see? Do you think the smooths were needed, to model this data?

ANSWER: The smooths were needed because the trends that the functions indicated were not linear. The smooths for mother_BMI & mother_height looked similar to each other, having a wave shape, while the plot for child_age was 'W' shaped. Because the predictors we chose are not linear, the smooths are needed to represent our predictors correctly.

- Processing. What do you think your results mean? If you were able to talk about them with parents in Zambia, or policymakers there, what would be important to communicate?

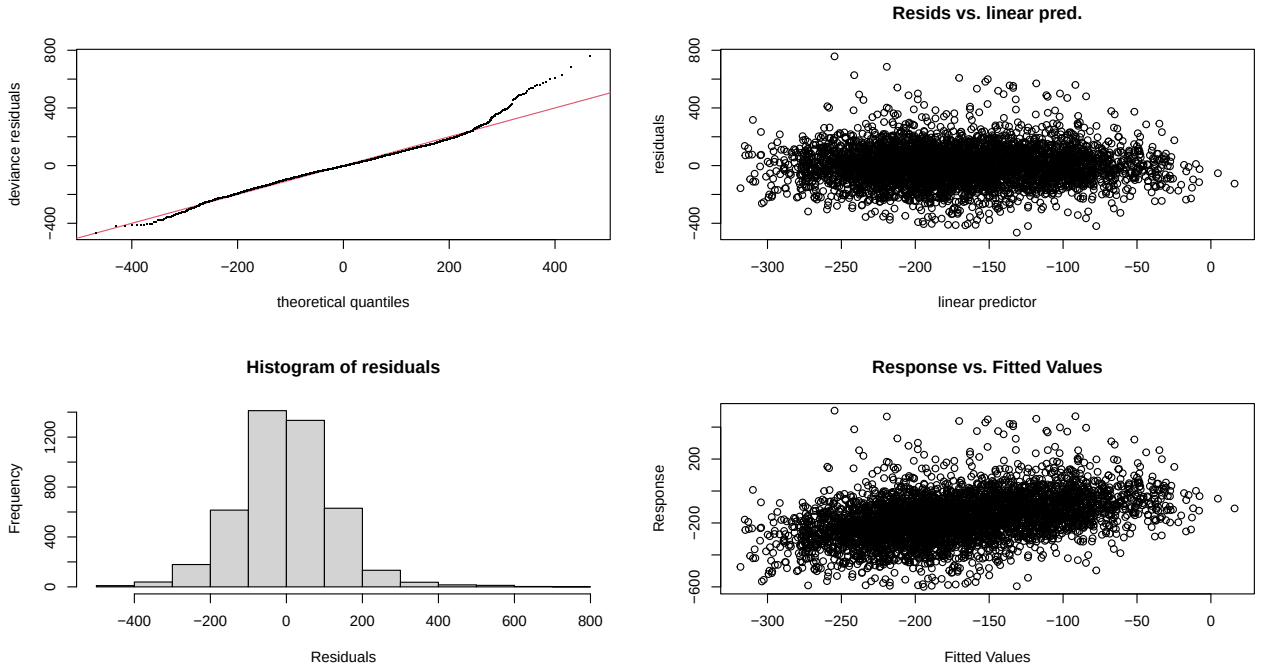
ANSWER: There is an optimal BMI that has the least z-score Between the age of 20 and 40, the child's predicted height z-score is the lowest There seems to be an optimal mother height that seems to be correlated with lower predicted height z-scores As mentioned in the background, these children do seem to be unusually small as a majority of the z-scores are well below zero

```
new.zam <- gam(height_zscore ~ s(child_age, k = 7, bs = 'cc') +
               s(mother_height, k = 7, bs = 'cc') +
                 mother_BMI,
               data = zam,
               method = 'ML',
               select = TRUE)
summary(new.zam)
```

```
##
## Family: gaussian
```

```
## Link function: identity
##
## Formula:
## height_zscore ~ s(child_age, k = 7, bs = "cc") + s(mother_height,
##     k = 7, bs = "cc") + mother_BMI
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -278.5200    12.7994  -21.76   <2e-16 ***
## mother_BMI     4.8790     0.5754    8.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F p-value
## s(child_age)     4.792      5 103.43  <2e-16 ***
## s(mother_height) 4.366      5  69.32  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.174   Deviance explained = 17.6%
## -ML =  27694  Scale est. = 16036     n = 4421
```

```
gam.check(new.zam)
```



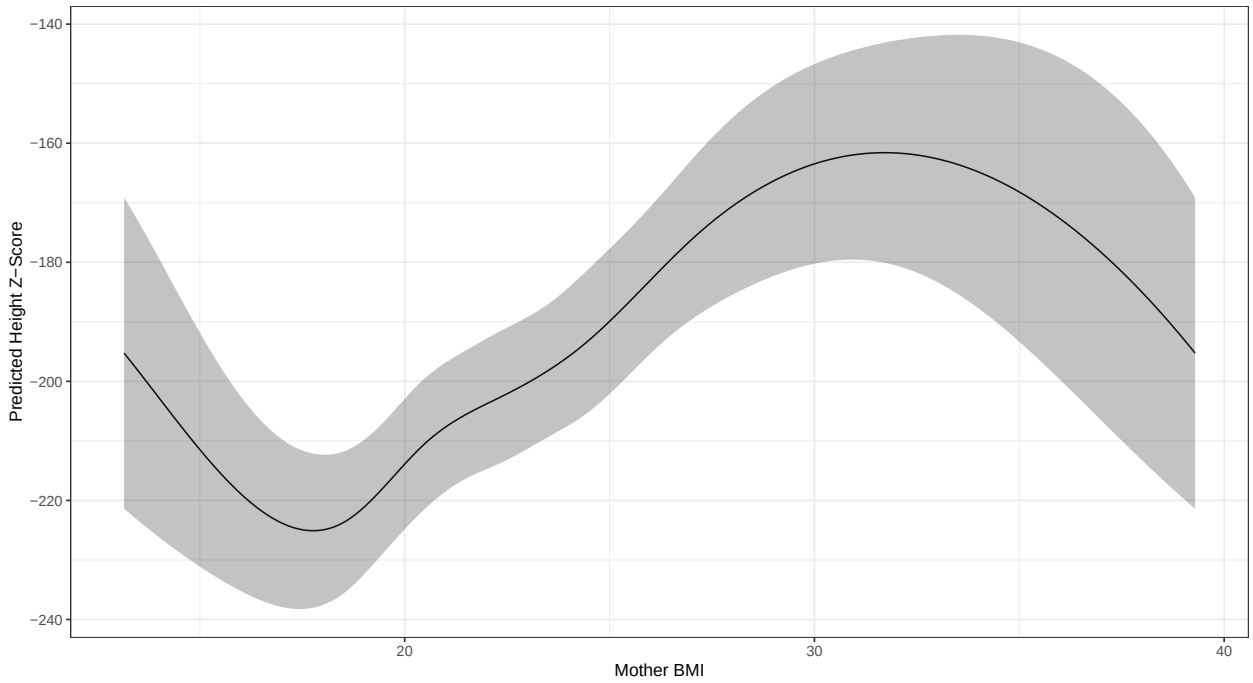

```
##
## Method: ML   Optimizer: outer newton
## full convergence after 7 iterations.
## Gradient range [-0.0001315307,0.0001209636]
## (score 27693.89 & scale 16035.6).
## Hessian positive definite, eigenvalue range [1.562173,2210.505].
## Model rank =  12 / 12
##
```

```
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                     k'  edf k-index p-value
## s(child_age)     5.00 4.79    0.92  <2e-16 ***
## s(mother_height) 5.00 4.37    0.90  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

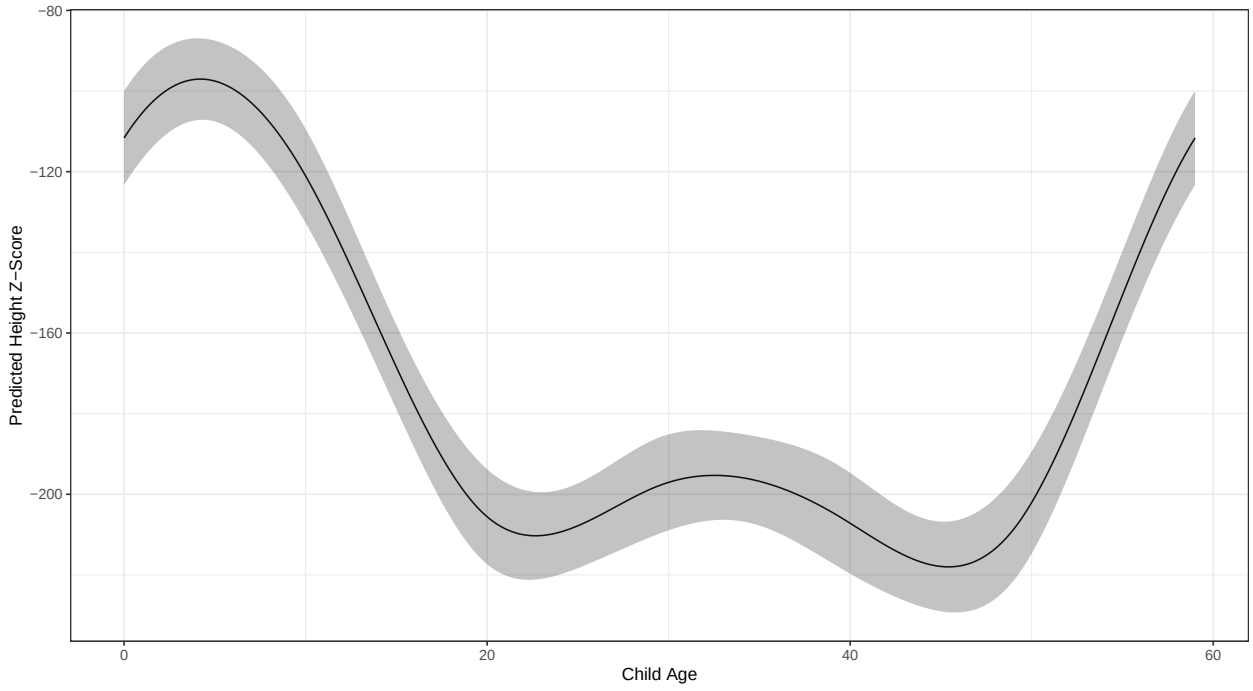
```
new.zam <- gam(height_zscore ~ s(child_age, k = 7, bs = 'cc') +
               s(mother_BMI, k = 7, bs = 'cc') +
               s(mother_height, k = 7, bs = 'cc'),
               data = zam,
               method = 'ML',
               select = TRUE)
summary(new.zam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## height_zscore ~ s(child_age, k = 7, bs = "cc") + s(mother_BMI,
##     k = 7, bs = "cc") + s(mother_height, k = 7, bs = "cc")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -171.190      1.905  -89.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                     edf Ref.df      F p-value
## s(child_age)     4.789      5 102.99  <2e-16 ***
## s(mother_BMI)    3.464      5  13.33  <2e-16 ***
## s(mother_height) 4.357      5  68.66  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.173   Deviance explained = 17.6%
## -ML =  27701  Scale est. = 16051     n = 4421
```

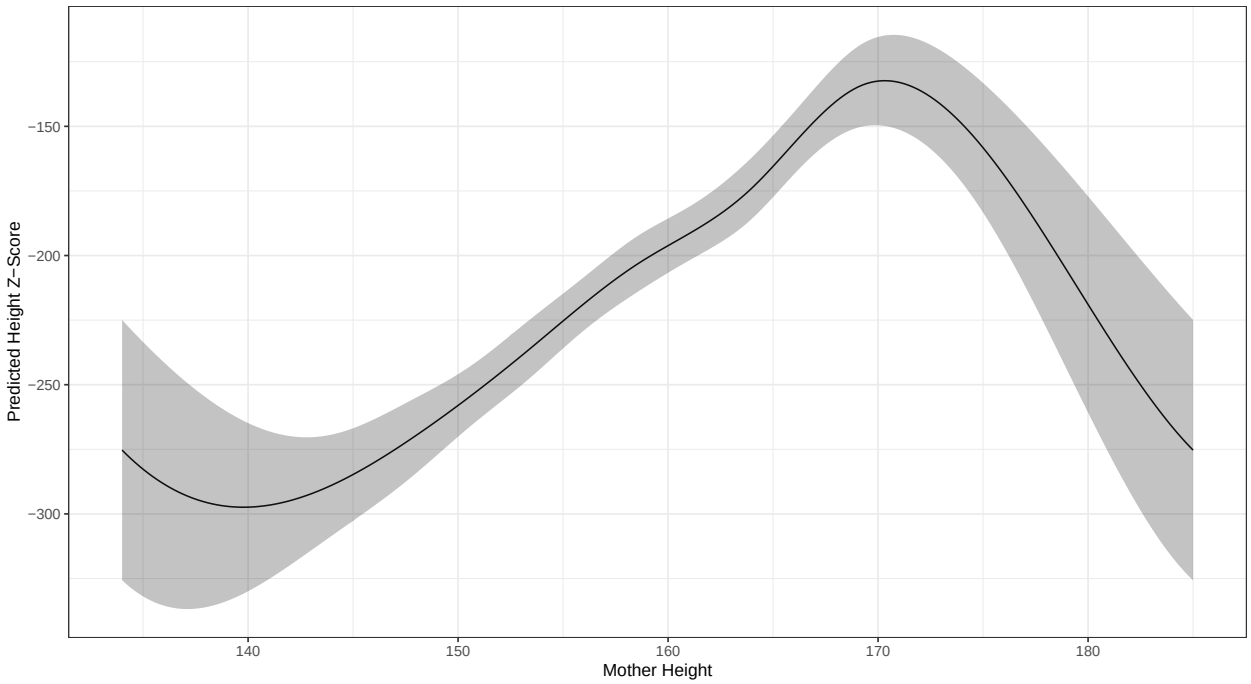
```
pred_plot(new.zam, 'mother_BMI') %>% gf_labs(y = 'Predicted Height Z-Score',
                                             x = 'Mother BMI')
```




```
pred_plot(new.zam, 'child_age') %>% gf_labs(y = 'Predicted Height Z-Score',
                                            x = 'Child Age')
```




```
pred_plot(new.zam, 'mother_height') %>% gf_labs(y = 'Predicted Height Z-Score',
                                                x = 'Mother Height')
```


Predicted Height Z−Score
−150
−200
−250
−300
140
150
160
170
180
Mother Height