Calculations for the Revisions

Alvin Sheng

2023-08-28

```
library(here)
```

```
## Warning in readLines(f, n): line 1 appears to contain an embedded nul
```

```
## Warning in readLines(f, n): incomplete final line found on
## '/Volumes/ALVINDRIVE2/flood-risk-health-effects/._flood-risk-health-effects.Rproj'
```

```
## here() starts at /Volumes/ALVINDRIVE2/flood-risk-health-effects
```

```
library(readxl)
library(stringr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v purrr     1.0.2
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(tidyr)
library(Matrix)
```

```
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(sf)
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

```
select <- dplyr::select
```

```
fhs_model_df <- readRDS(here("intermediary_data/fhs_model_df_fr_and_pollute_pc.rds"))
```

Double checking summary statistics of missingness across the responses and covariates

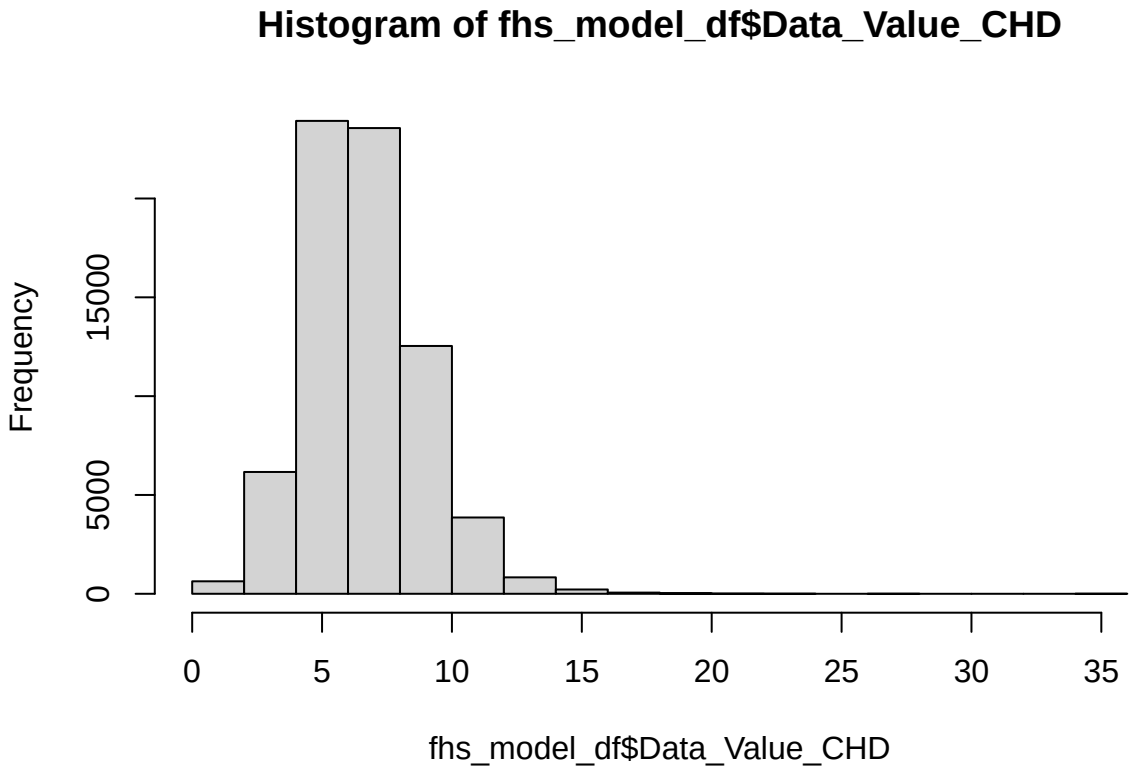
In particular, the median and maximum

```
summary(apply(fhs_model_df[, 19:ncol(fhs_model_df)], 2, function(vec) {mean(is.na(vec))}))
```

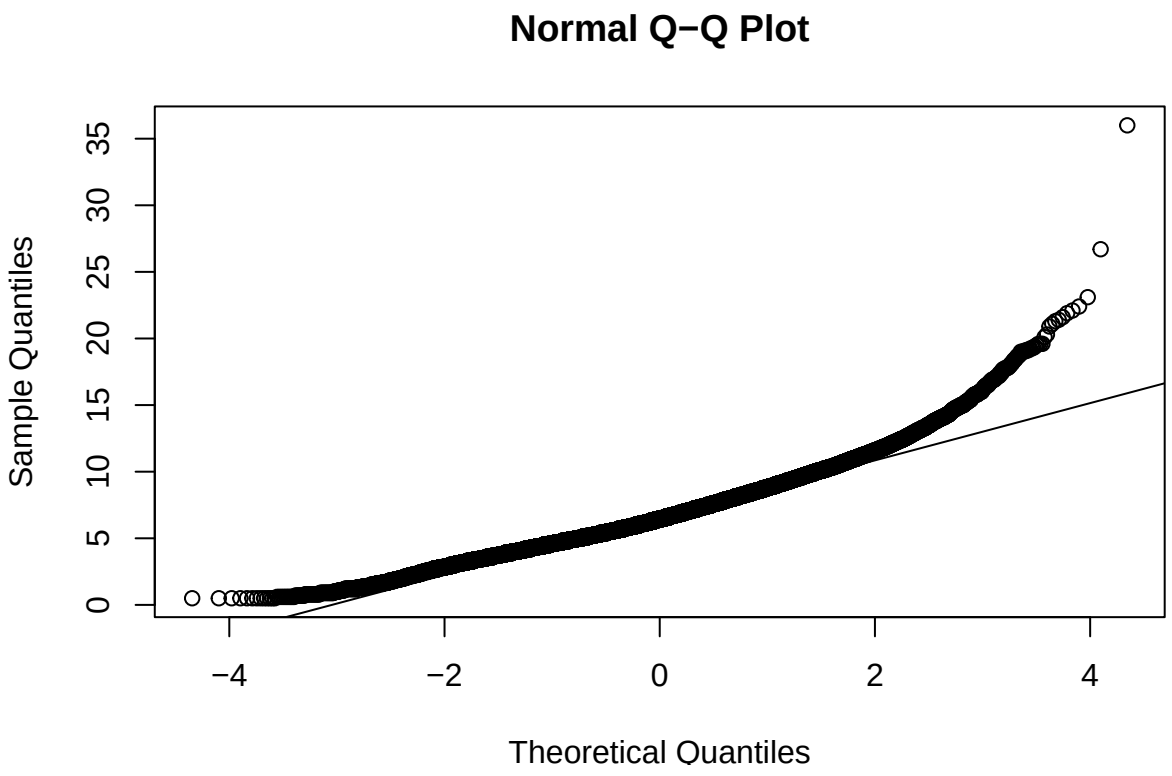
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0003171 0.0030329 0.0096778 0.0176973 0.0105257 0.0735625
```

Checking distribution of the responses

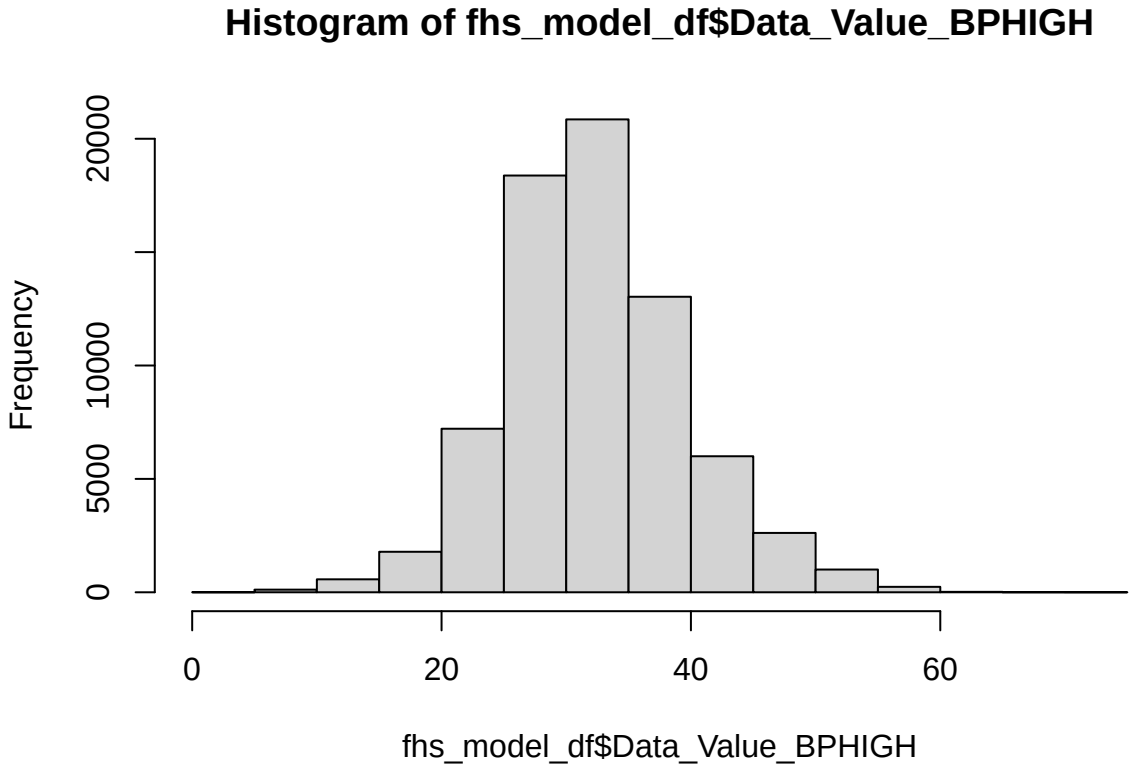
```
hist(fhs_model_df$Data_Value_CHD)
```



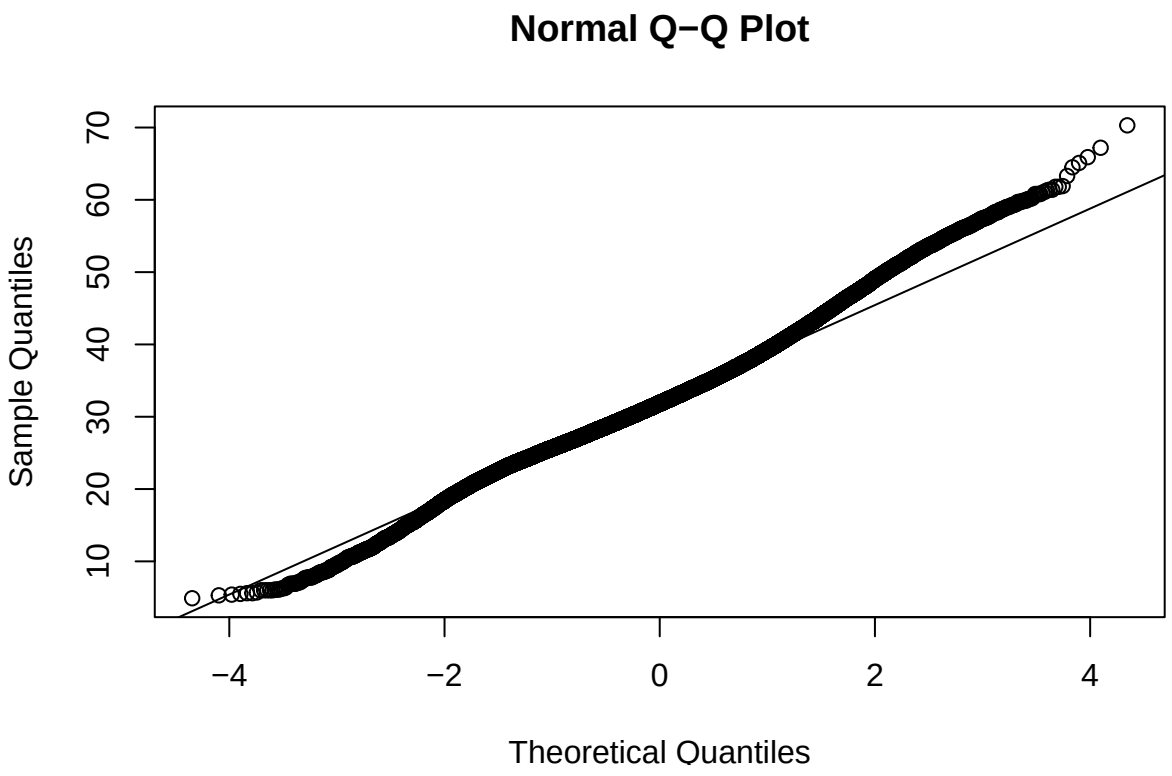

```
qqnorm(fhs_model_df$Data_Value_CHD)
qqline(fhs_model_df$Data_Value_CHD)
```



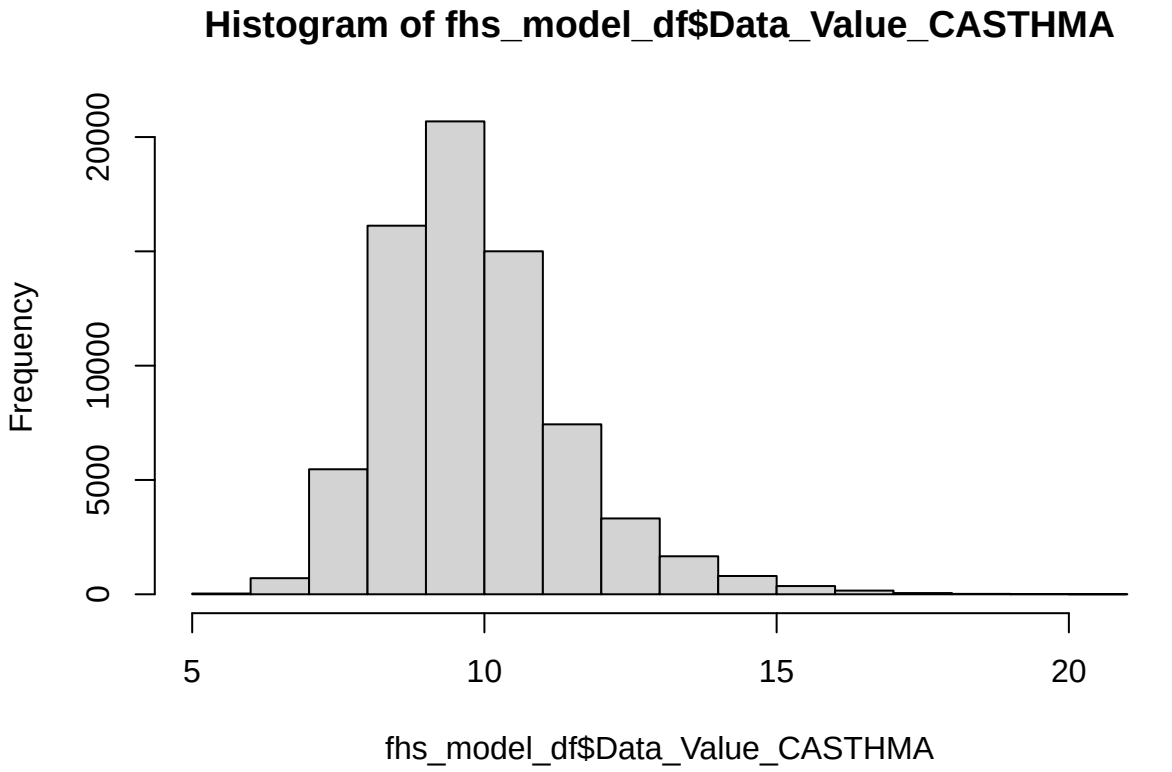

```
hist(fhs_model_df$Data_Value_BPHIGH)
```



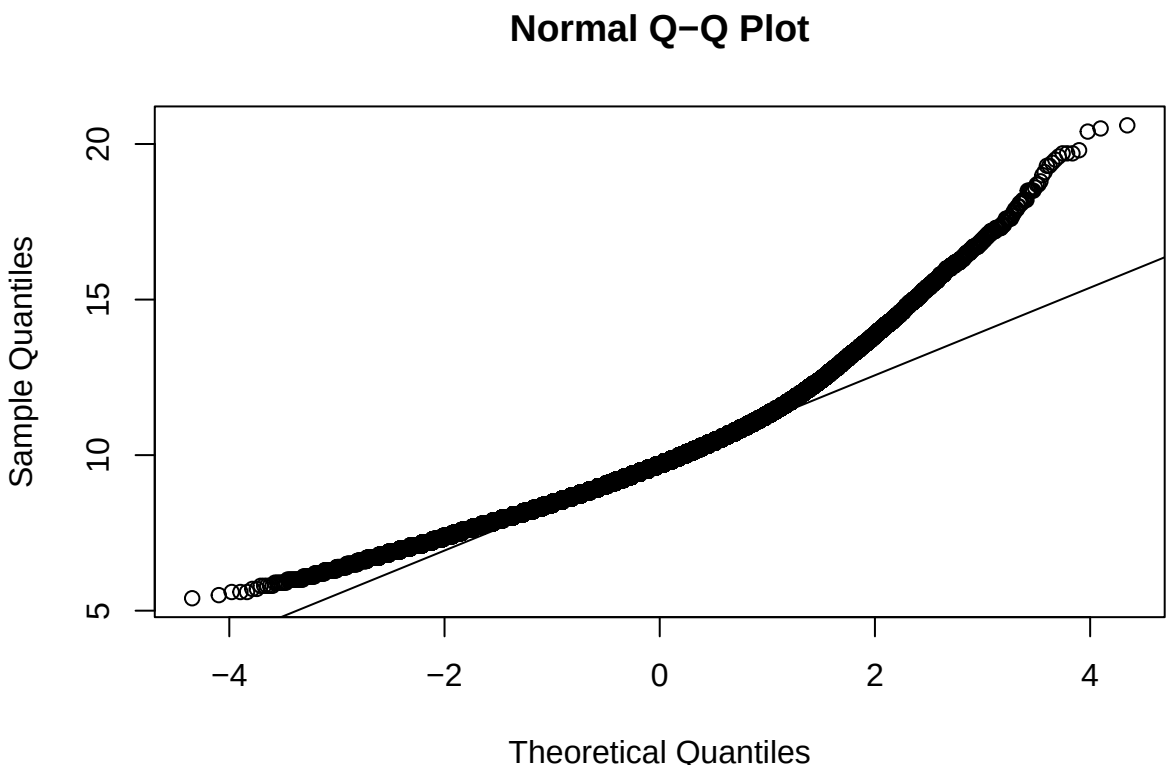

```
qqnorm(fhs_model_df$Data_Value_BPHIGH)
qqline(fhs_model_df$Data_Value_BPHIGH)
```



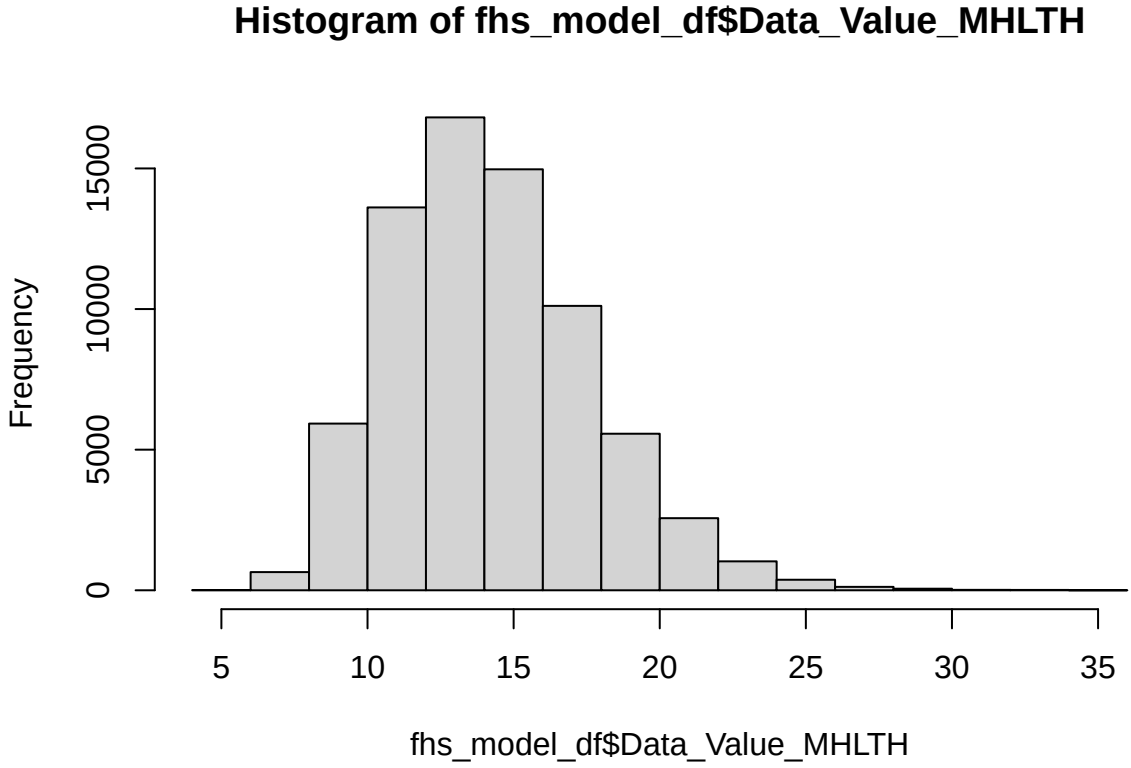

```
hist(fhs_model_df$Data_Value_CASTHMA)
```



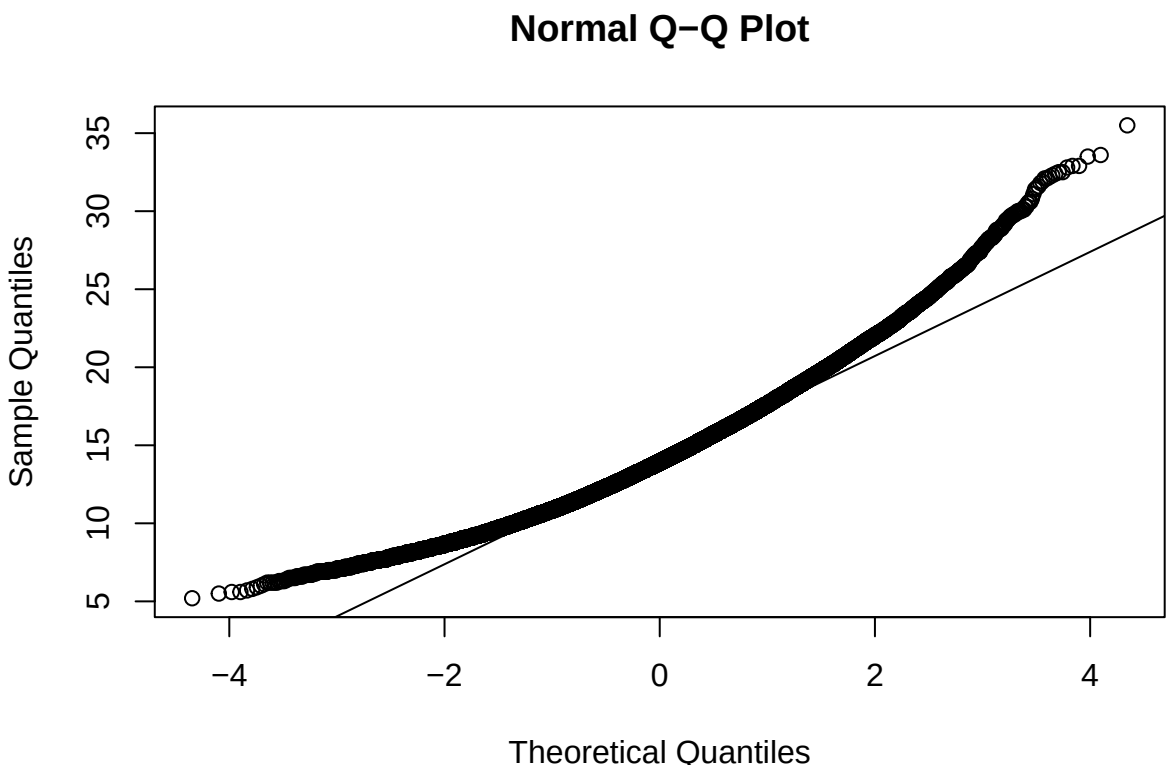

```
qqnorm(fhs_model_df$Data_Value_CASTHMA)
qqline(fhs_model_df$Data_Value_CASTHMA)
```




```
hist(fhs_model_df$Data_Value_MHLTH)
```




```
qqnorm(fhs_model_df$Data_Value_MHLTH)
qqline(fhs_model_df$Data_Value_MHLTH)
```



Variability of response variables among census tracts

```
# work with the PLACES dataset
# https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/cwsq

places_dat <- read.csv(here("imported_data",
                            "PLACES__Local_Data_for_Better_Health__Census_Tract_Data_2020_release.csv"))

# don't need the year, state_abbr, lat or lon

places_subset <- dplyr::select(places_dat, -c(Year, StateAbbr, StateDesc, LocationName,
                                              DataSource, Category, Data_Value_Unit, Data_Value_Type,
                                              Data_Value_Footnote_Symbol, Data_Value_Footnote,
                                              Geolocation, DataValueTypeID))

# convert from long to wide format

places_dat_wide <- pivot_wider(places_subset, id_cols = c(LocationID, CountyFIPS, TotalPopulation),
                               names_from = MeasureId,
                               values_from = c(Data_Value, Low_Confidence_Limit, High_Confidence_Limit))

places_dat_wide <- rename(places_dat_wide, fips = LocationID)
```

```
# checking heteroskedasticity
summary(places_dat_wide$High_Confidence_Limit_CHD - places_dat_wide$Low_Confidence_Limit_CHD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.800   1.000   1.086   1.300   9.800
```

```
sd(places_dat_wide$Data_Value_CHD)
```

```
## [1] 2.205924
```

```
summary(places_dat_wide$High_Confidence_Limit_BPHIGH - places_dat_wide$Low_Confidence_Limit_BPHIGH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   1.700   2.000   2.113   2.500   9.300
```

```
sd(places_dat_wide$Data_Value_BPHIGH)
```

```
## [1] 7.286122
```

```
summary(places_dat_wide$High_Confidence_Limit_CASTHMA - places_dat_wide$Low_Confidence_Limit_CASTHMA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.300   0.800   1.000   1.058   1.200   5.700
```

```
sd(places_dat_wide$Data_Value_CASTHMA)
```

```
## [1] 1.574586
```

```
summary(places_dat_wide$High_Confidence_Limit_MHLTH - places_dat_wide$Low_Confidence_Limit_MHLTH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.50    2.00    2.50    2.67    3.10   16.40
```

```
sd(places_dat_wide$Data_Value_MHLTH)
```

```
## [1] 3.408443
```

Checking Census Tract Statistics

```
summary(fhs_model_df$E_TOTPOP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    2924    4128    4434    5542   70271     220
```

```
summary(fhs_model_df$E_HU)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1275    1753    1874    2333   26436     220
```

```
summary(fhs_model_df$E_HH)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0    1105    1544    1645    2065   21337     220
```

```
sd(fhs_model_df$E_TOTPOP, na.rm = T)
```

```
## [1] 2275.316
```

```
sd(fhs_model_df$E_HU, na.rm = T)
```

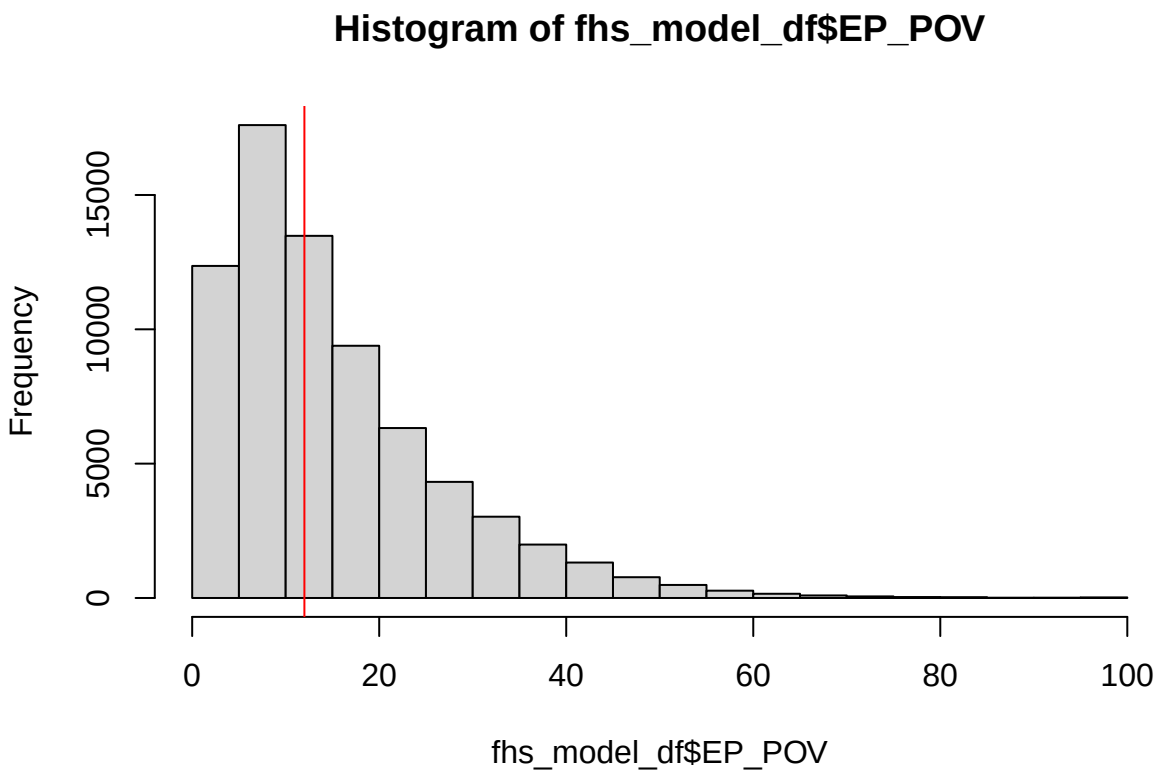
```
## [1] 910.691
```

```
sd(fhs_model_df$E_HH, na.rm = T)
```

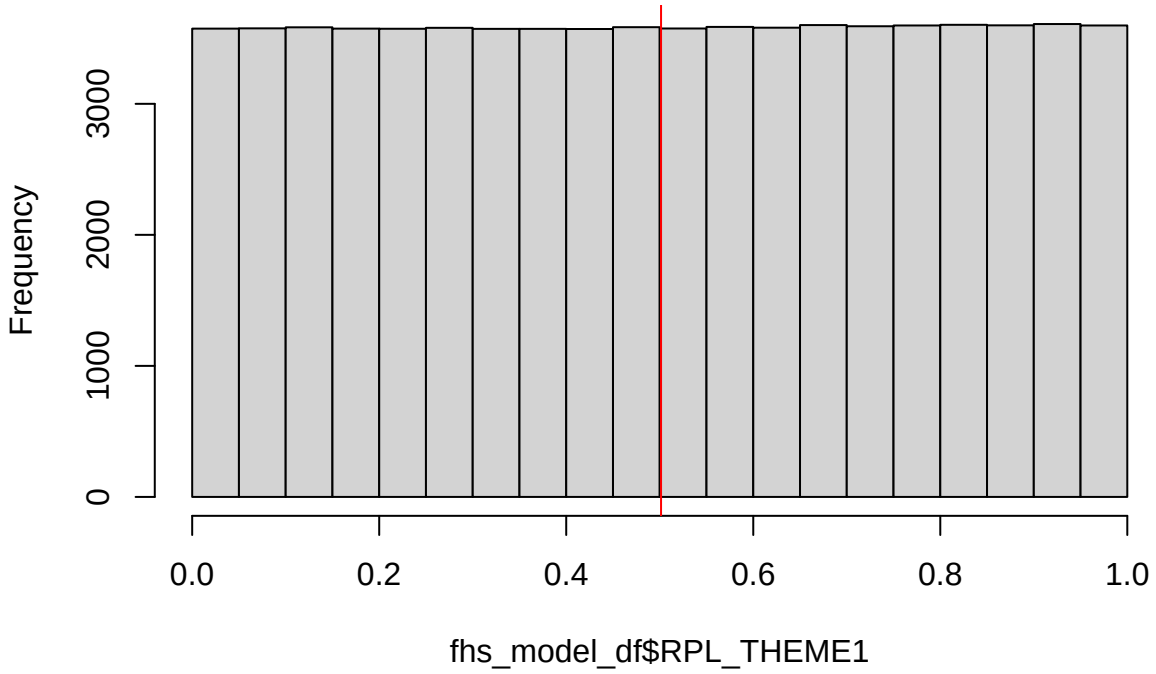
```
## [1] 809.4712
```

Checking distributions of the stratification variables

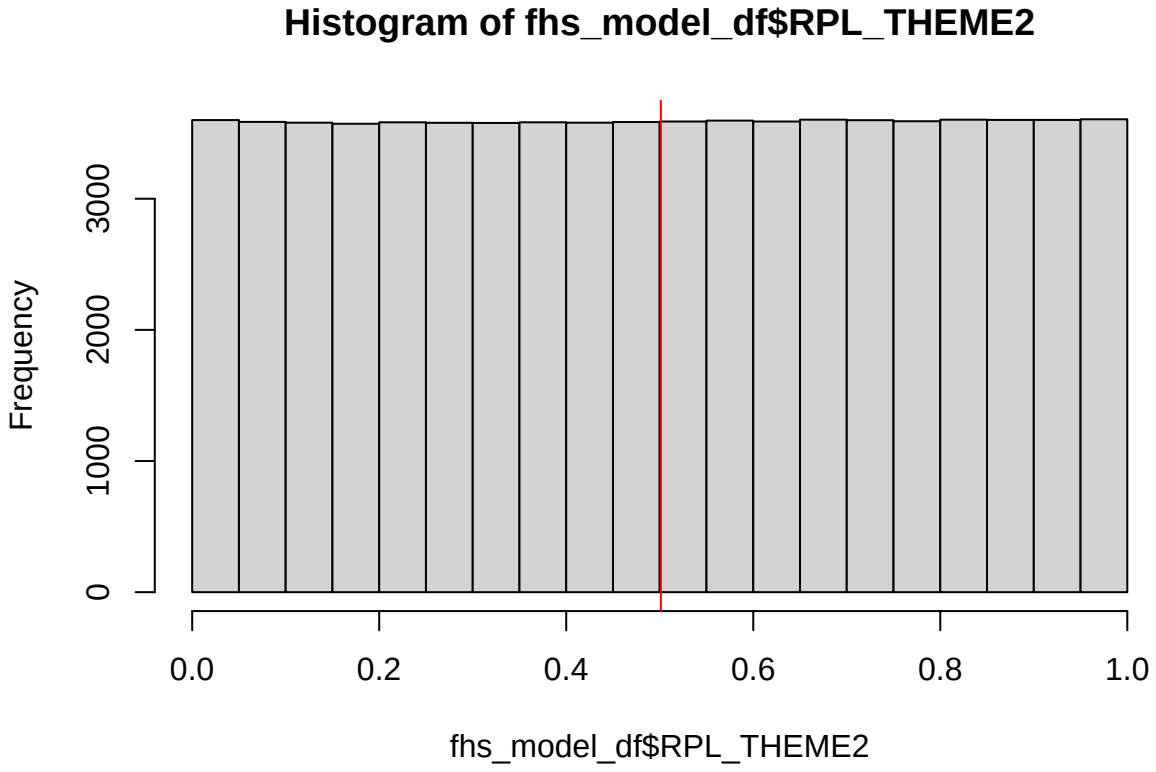
```
hist(fhs_model_df$EP_POV)
abline(v = median(fhs_model_df$EP_POV, na.rm = T), col = "red")
```



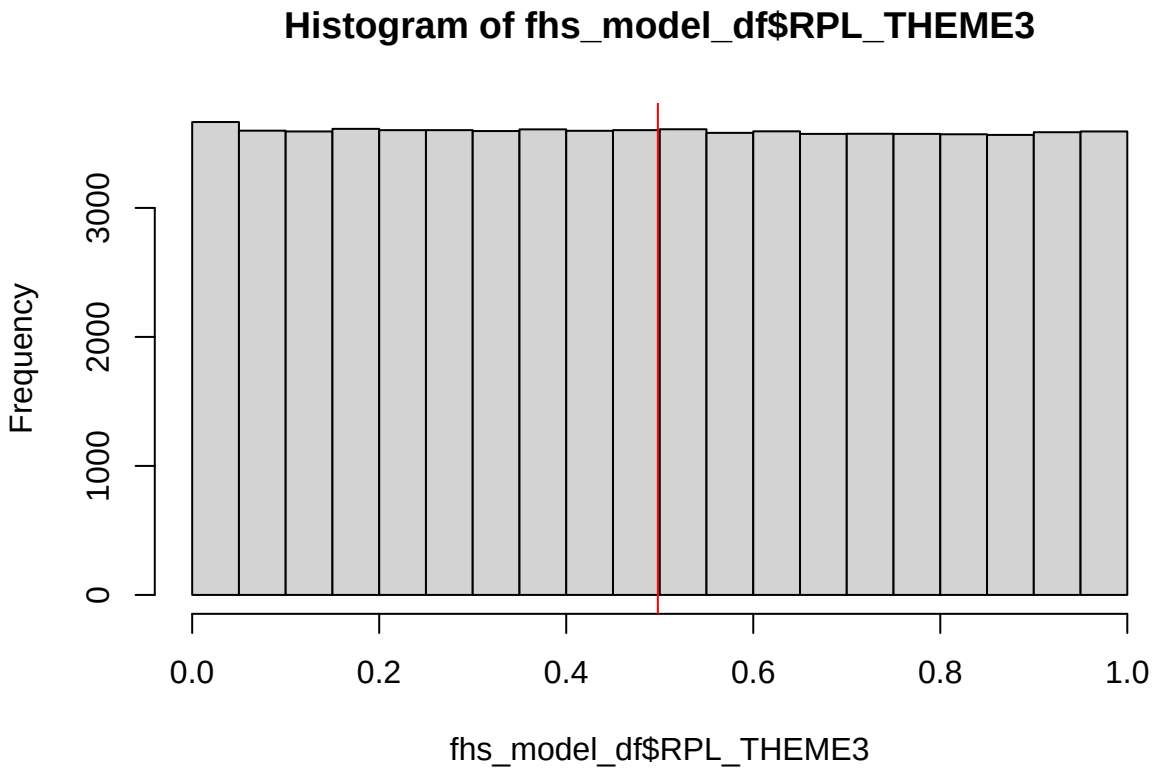

```
hist(fhs_model_df$RPL_THEME1)
abline(v = median(fhs_model_df$RPL_THEME1, na.rm = T), col = "red")
```



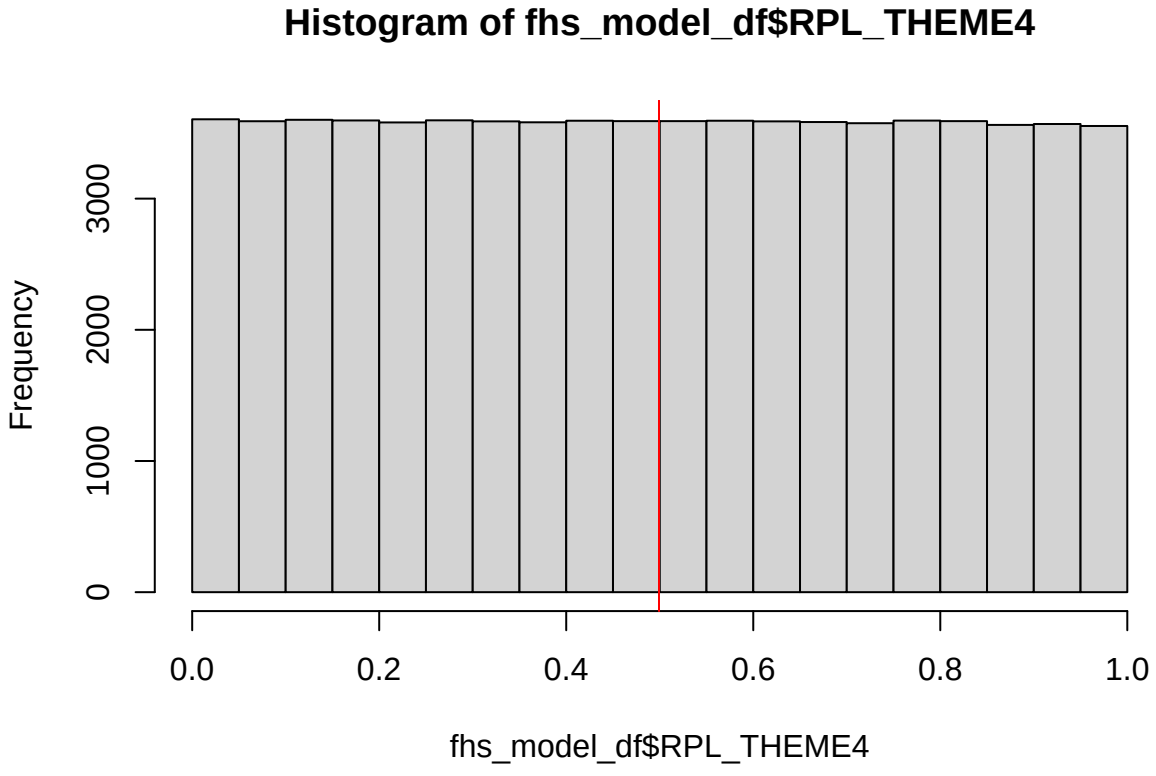

```
hist(fhs_model_df$RPL_THEME2)
abline(v = median(fhs_model_df$RPL_THEME2, na.rm = T), col = "red")
```



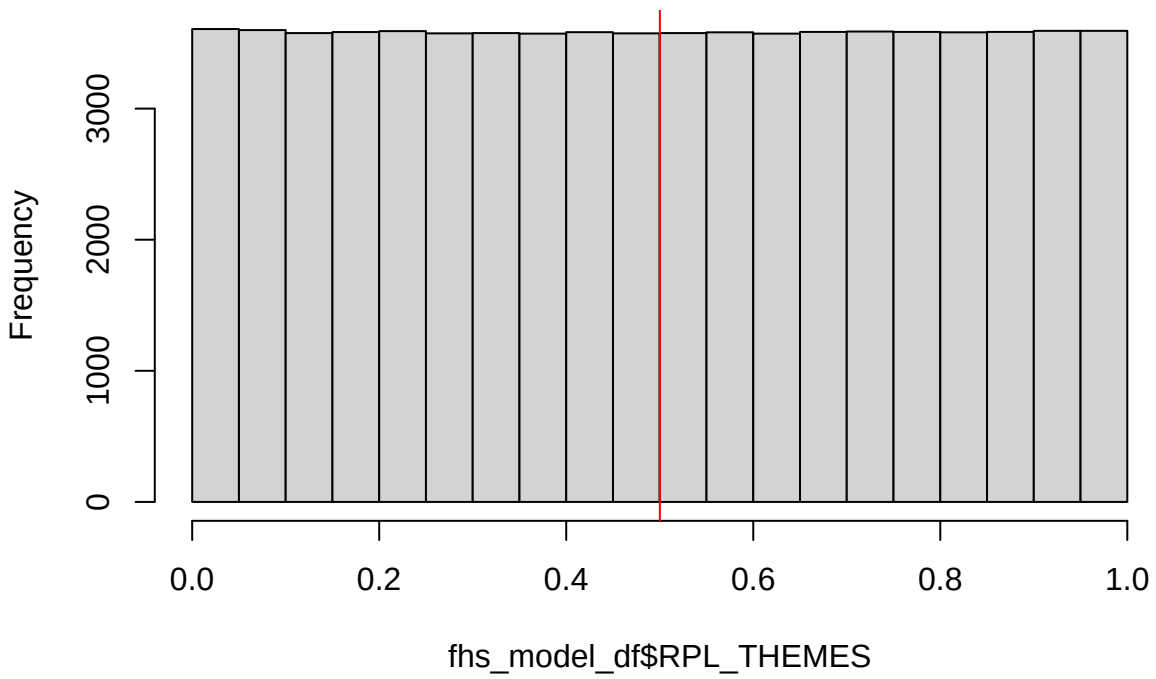

```
hist(fhs_model_df$RPL_THEME3)
abline(v = median(fhs_model_df$RPL_THEME3, na.rm = T), col = "red")
```




```
hist(fhs_model_df$RPL_THEME4)
abline(v = median(fhs_model_df$RPL_THEME4, na.rm = T), col = "red")
```




```
hist(fhs_model_df$RPL_THEMES)
abline(v = median(fhs_model_df$RPL_THEMES, na.rm = T), col = "red")
```

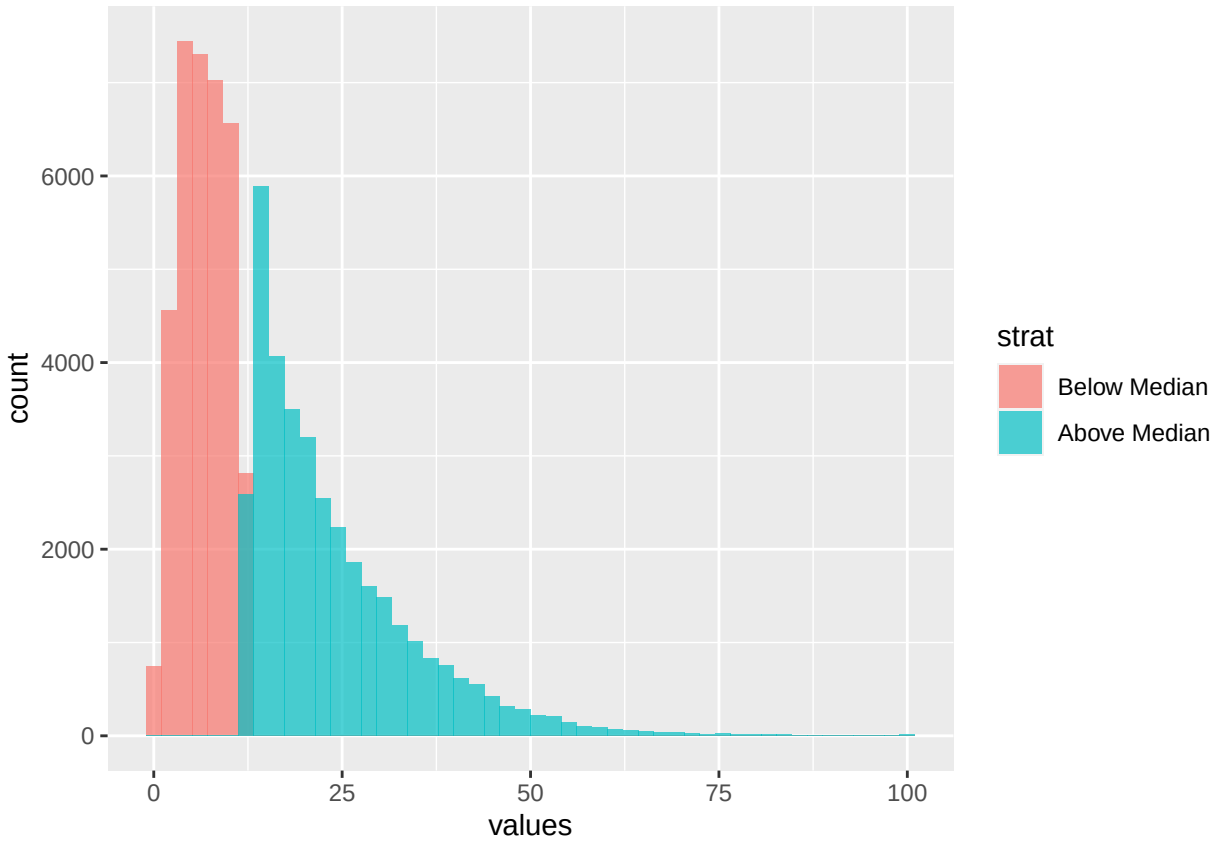


Doing single histogram with median line, for single SVIs

```
par(mfrow = c(5, 3))

hist(fhs_model_df$EP_POV, main = "% Below Poverty Level", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_POV, na.rm = T), col = "red")

hist(fhs_model_df$EP_UNEMP, main = "% Unemployed", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_UNEMP, na.rm = T), col = "red")

hist(fhs_model_df$EP_PCI, main = "Per Capita Income", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_PCI, na.rm = T), col = "red")

hist(fhs_model_df$EP_NOHSDP, main = "% No High School Diploma", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_NOHSDP, na.rm = T), col = "red")

hist(fhs_model_df$EP_AGE65, main = "% Over 65", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_AGE65, na.rm = T), col = "red")

hist(fhs_model_df$EP_AGE17, main = "% Under 17", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_AGE17, na.rm = T), col = "red")

hist(fhs_model_df$EP_DISABL, main = "% Disabled", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_DISABL, na.rm = T), col = "red")

hist(fhs_model_df$EP_SNGPNT, main = "% Households Single Parent", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_SNGPNT, na.rm = T), col = "red")

hist(fhs_model_df$EP_MINRTY, main = "% Minority", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_MINRTY, na.rm = T), col = "red")
```

```r
hist(fhs_model_df$EP_LIMENG, main = "% Limited English", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_LIMENG, na.rm = T), col = "red")

hist(fhs_model_df$EP_MUNIT, main = "% Housing With More Than 10 Units", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_MUNIT, na.rm = T), col = "red")

hist(fhs_model_df$EP_MOBILE, main = "% Mobile Homes", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_MOBILE, na.rm = T), col = "red")

hist(fhs_model_df$EP_CROWD, main = "% Crowded Housing", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_CROWD, na.rm = T), col = "red")

hist(fhs_model_df$EP_NOVEH, main = "% Households Without Vehicles", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_NOVEH, na.rm = T), col = "red")

hist(fhs_model_df$EP_GROUPQ, main = "% Persons in Group Quarters", xlab = NULL, ylab= NULL)
abline(v = median(fhs_model_df$EP_GROUPQ, na.rm = T), col = "red")

# export as US Letter, Portrait, as figures/final_figures/SVI_distributions.pdf
```

```r
strat_fn <- median

strat_covariate <- fhs_model_df$EP_POV
strat_covariate[is.na(strat_covariate)] <- mean(strat_covariate, na.rm = T)

strat <- ifelse(strat_covariate <= strat_fn(strat_covariate), "Below Median", "Above Median")
strat <- factor(strat, levels = c("Below Median", "Above Median"))

hist_dat <- data.frame(values = strat_covariate, strat = strat)

ggplot(hist_dat, aes(x = values, fill = strat)) +                        # Draw overlaying histogram
  geom_histogram(position = "identity", alpha = 0.7, bins = 50)
```



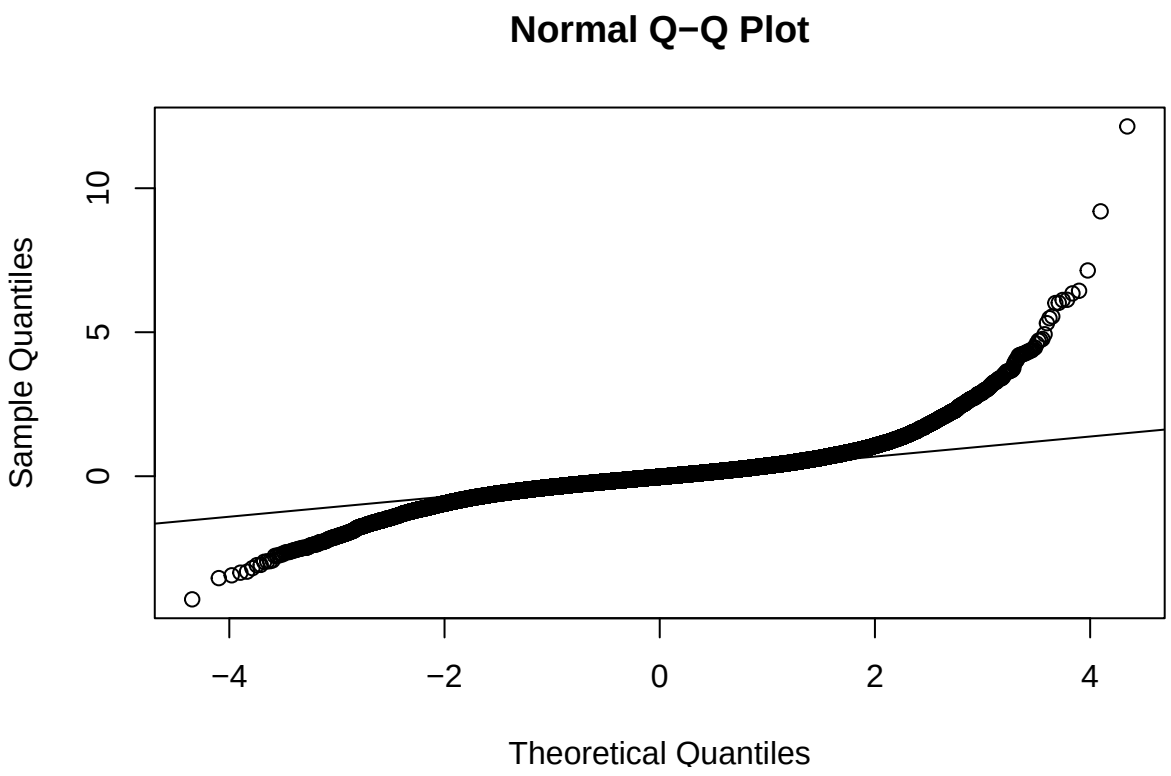
Checking Linear Regression Assumptions for Models Stratified on all SVI Themes

```
# loading requisite data

load(here("modeling_files/stratified_analysis/model_stratif_rpls.RData"))
CHD.mean.fitted <- (chain1$mean.fitted + chain2$mean.fitted + chain3$mean.fitted) / 3
CHD.mean.phi <- (chain1$samples$mean.phi + chain2$samples$mean.phi + chain3$samples$mean.phi) / 3
CHD.mean.resid <- (chain1$residuals$response + chain2$residuals$response + chain3$residuals$response) /
CHD.mean.beta <- apply(rbind(chain1$samples$beta, chain2$samples$beta, chain3$samples$beta), 2, mean)

load(here("modeling_files/stratified_analysis/model_stratif_rpls_BPHIGH.RData"))
BPHIGH.mean.fitted <- (chain1$mean.fitted + chain2$mean.fitted + chain3$mean.fitted) / 3
BPHIGH.mean.phi <- (chain1$samples$mean.phi + chain2$samples$mean.phi + chain3$samples$mean.phi) / 3
BPHIGH.mean.resid <- (chain1$residuals$response + chain2$residuals$response + chain3$residuals$response
BPHIGH.mean.beta <- apply(rbind(chain1$samples$beta, chain2$samples$beta, chain3$samples$beta), 2, mean

load(here("modeling_files/stratified_analysis/model_stratif_rpls_CASTHMA.RData"))
CASTHMA.mean.fitted <- (chain1$mean.fitted + chain2$mean.fitted + chain3$mean.fitted) / 3
CASTHMA.mean.phi <- (chain1$samples$mean.phi + chain2$samples$mean.phi + chain3$samples$mean.phi) / 3
CASTHMA.mean.resid <- (chain1$residuals$response + chain2$residuals$response + chain3$residuals$respons
CASTHMA.mean.beta <- apply(rbind(chain1$samples$beta, chain2$samples$beta, chain3$samples$beta), 2, mea

load(here("modeling_files/stratified_analysis/model_stratif_rpls_MHLTH.RData"))
```

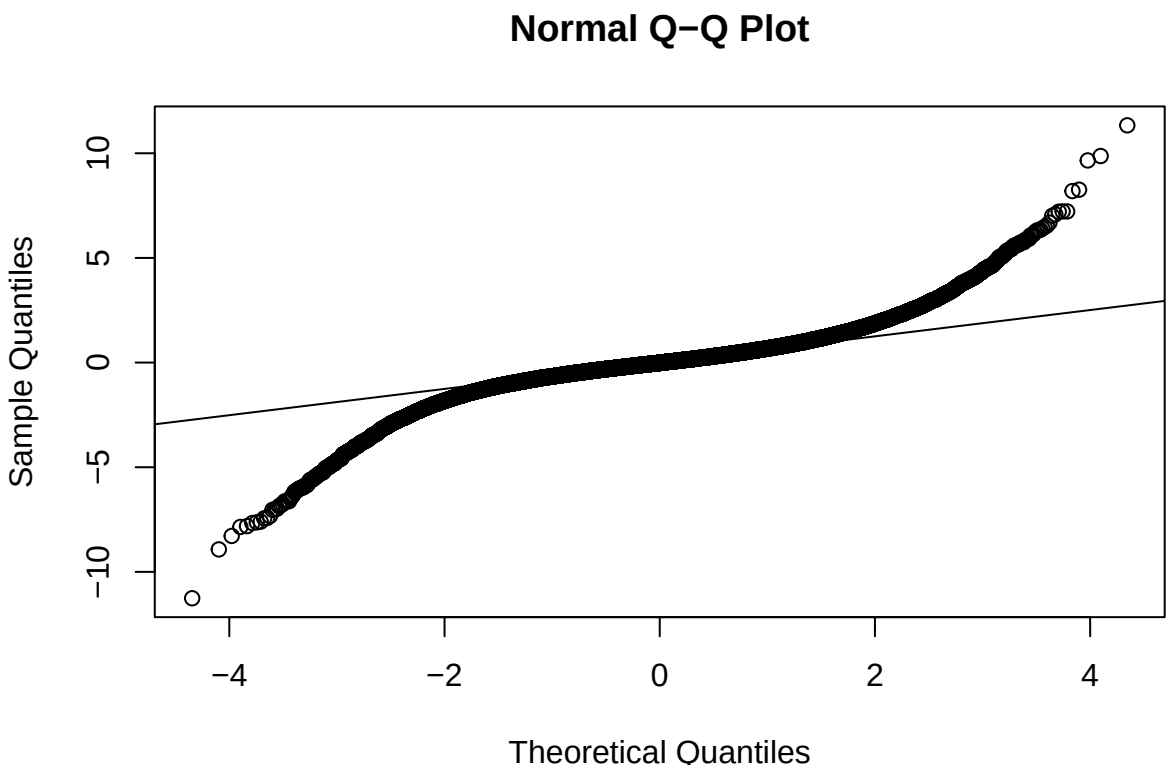
```
MHLTH.mean.fitted <- (chain1$mean.fitted + chain2$mean.fitted + chain3$mean.fitted) / 3
MHLTH.mean.phi <- (chain1$samples$mean.phi + chain2$samples$mean.phi + chain3$samples$mean.phi) / 3
MHLTH.mean.resid <- (chain1$residuals$response + chain2$residuals$response + chain3$residuals$response)
MHLTH.mean.beta <- apply(rbind(chain1$samples$beta, chain2$samples$beta, chain3$samples$beta), 2, mean)
```

QQ-plots

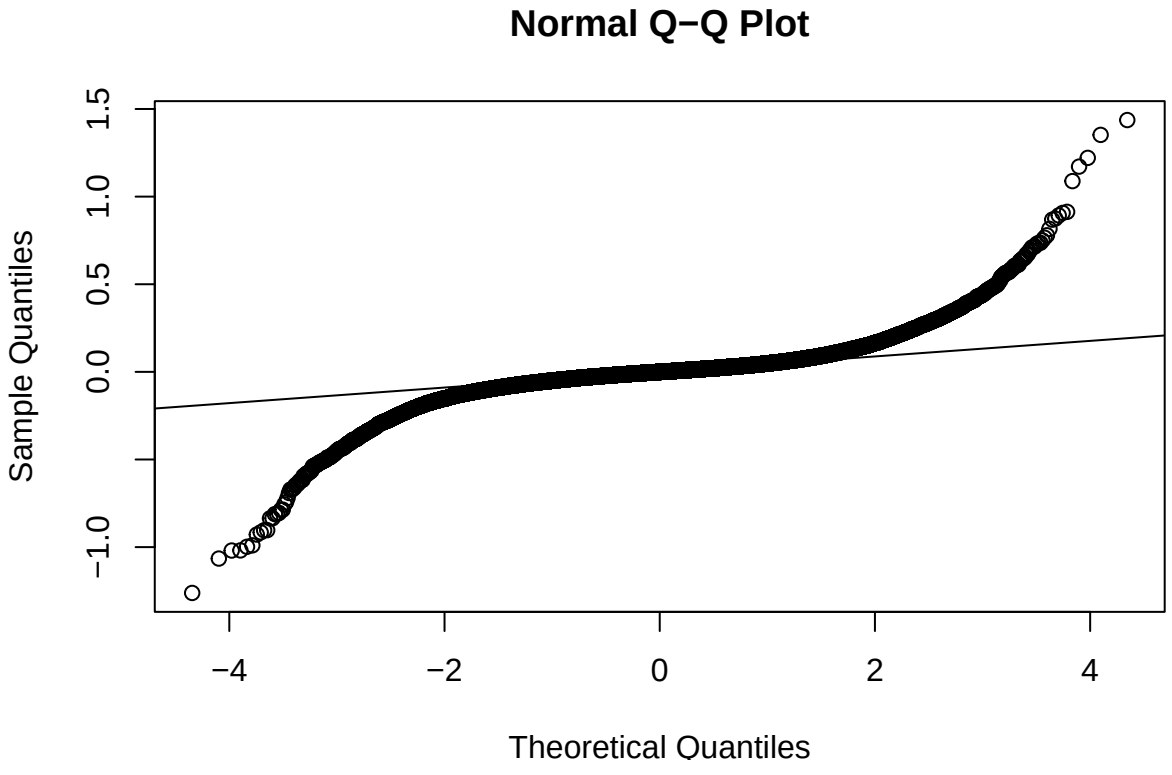
```
qqnorm(CHD.mean.resid)
qqline(CHD.mean.resid)
```



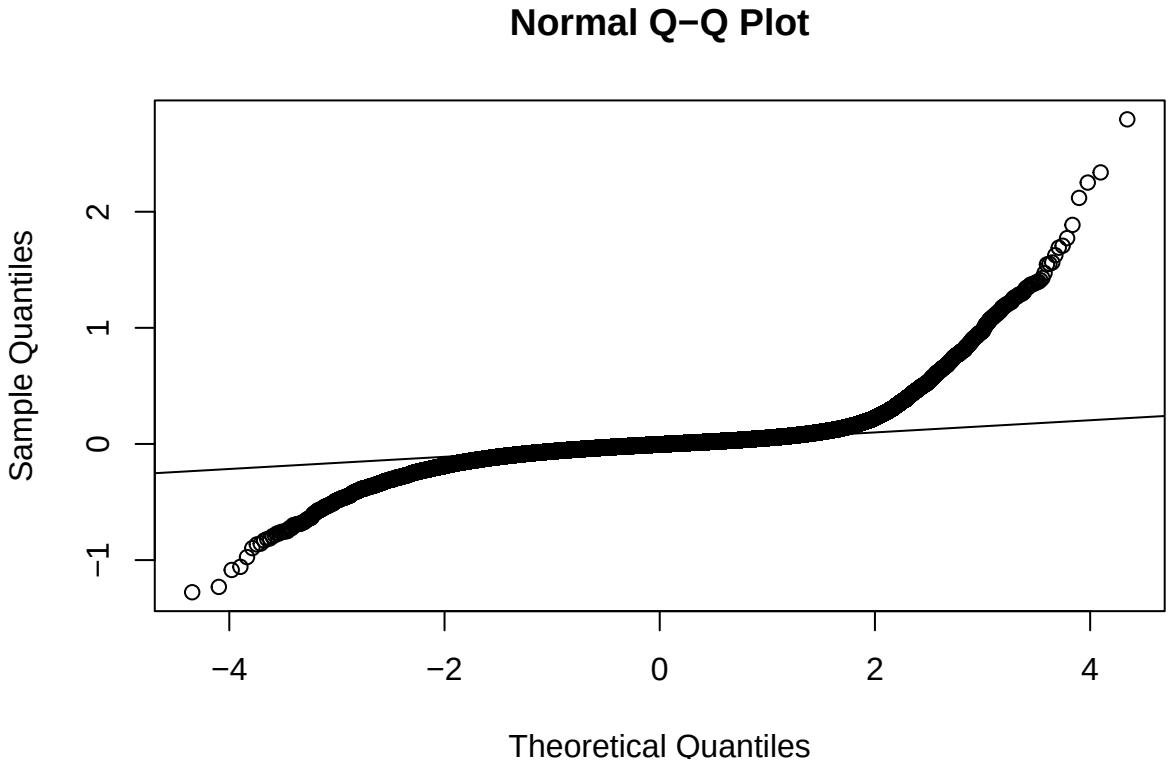

```
qqnorm(BPHIGH.mean.resid)
qqline(BPHIGH.mean.resid)
```




```
qqnorm(CASTHMA.mean.resid)
qqline(CASTHMA.mean.resid)
```

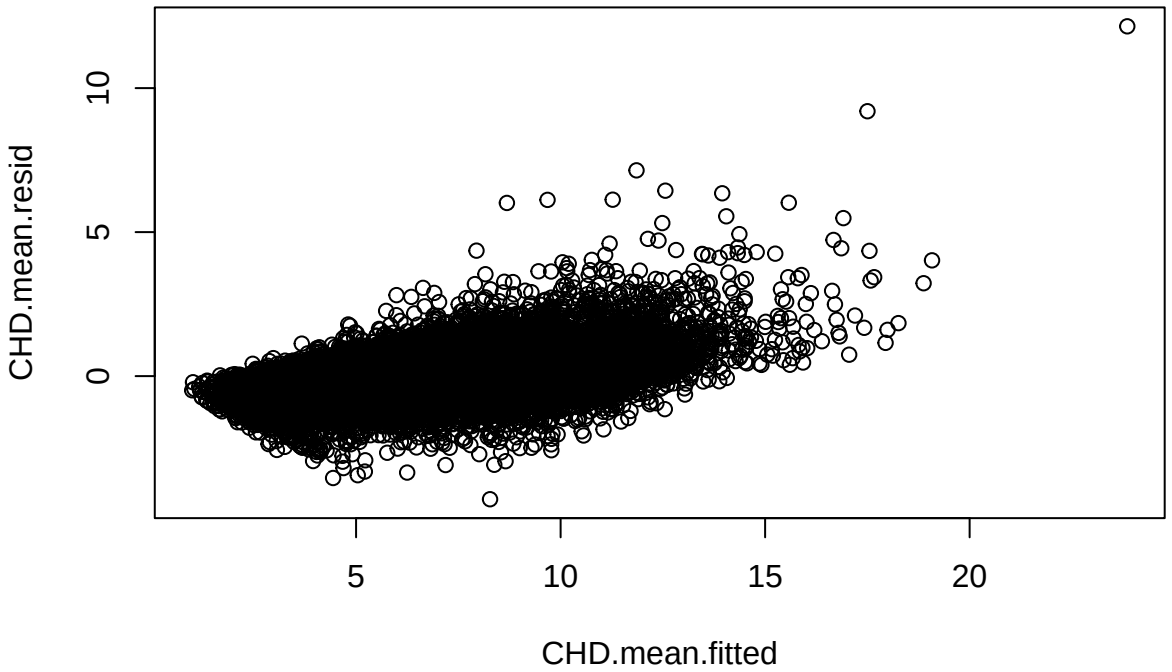



```
qqnorm(MHLTH.mean.resid)
qqline(MHLTH.mean.resid)
```

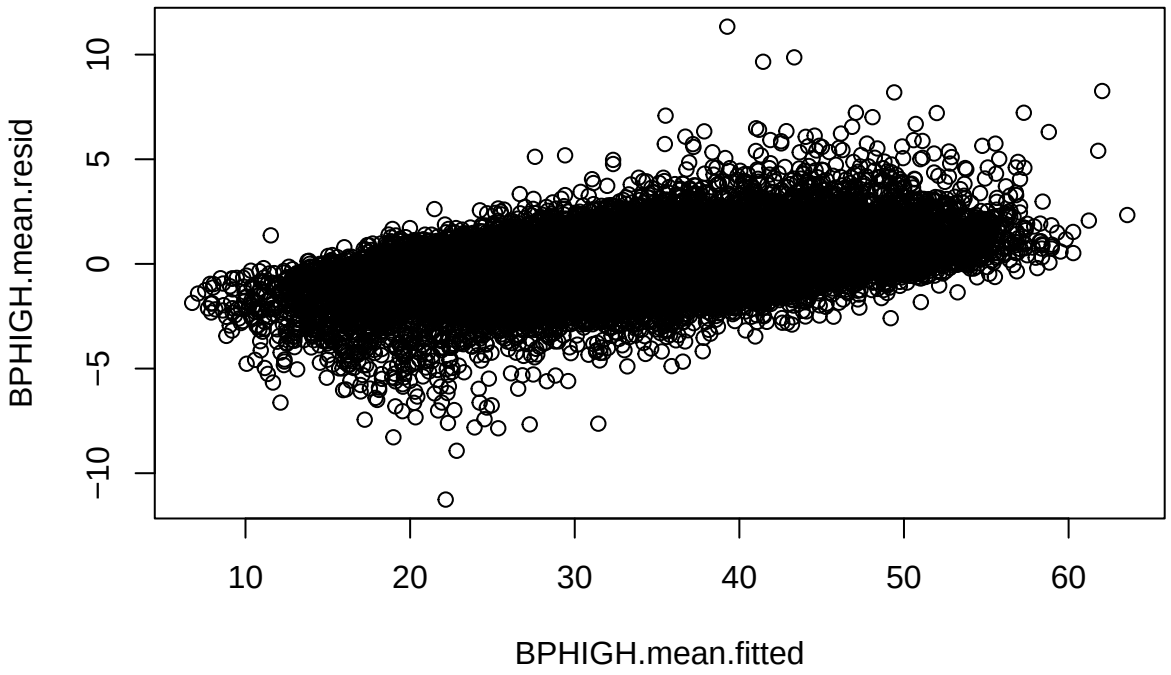


Residuals vs. Fitted

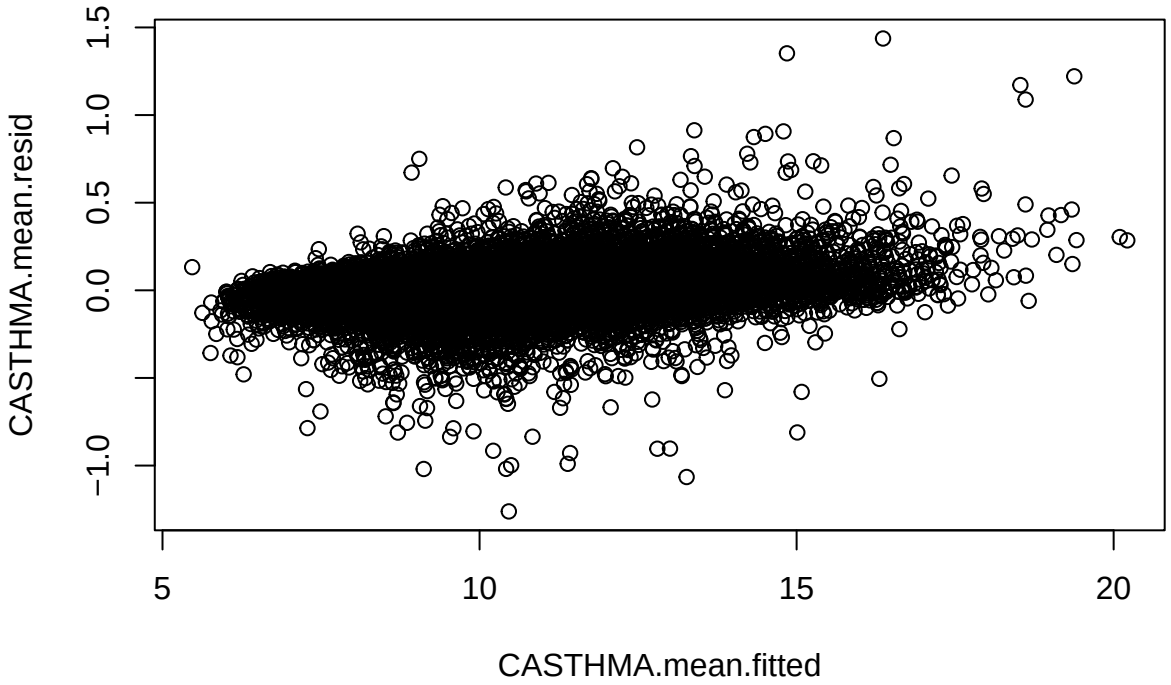
```
plot(CHD.mean.fitted, CHD.mean.resid)
```



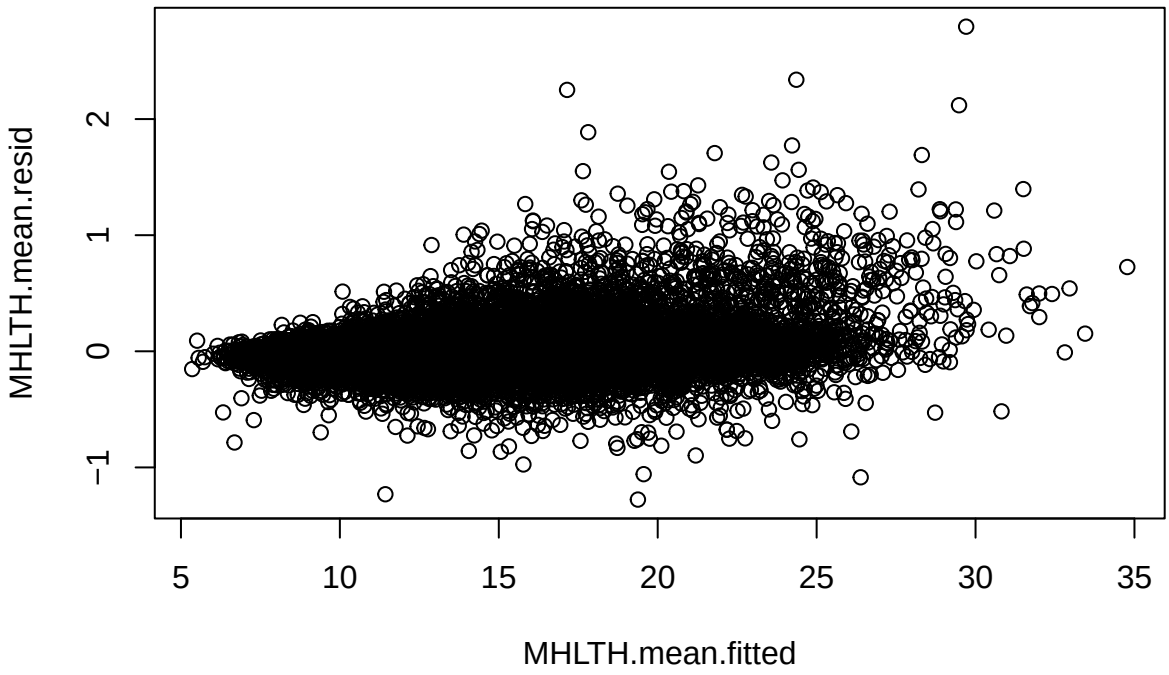

```
plot(BPHIGH.mean.fitted, BPHIGH.mean.resid)
```




```
plot(CASTHMA.mean.fitted, CASTHMA.mean.resid)
```




```
plot(MHLTH.mean.fitted, MHLTH.mean.resid)
```


MHLTH.mean.resid
−1
0
1
2
5
10
15
20
25
30
35
MHLTH.mean.fitted