Prelim Models

2025-02-26

Load packages

```
library(tidyverse)
library(readr)
library(readxl)
library(splines)
library(mgcv)
```

Template

```
knitr::opts_chunk$set(
  fig.width = 6,
  fig.asp = .6,
  out.width = "90%"
)

theme_set(theme_minimal() + theme(legend.position = "right"))

options(
  ggplot2.continuous.colour = "viridis",
  ggplot2.continuous.fill = "viridis"
)

scale_colour_discrete = scale_colour_viridis_d
scale_fill_discrete = scale_fill_viridis_d
```

#Load dataset Clean dataset has unreliable low bw values and the NA low bw values removed

```
eqi_lbw_clean_df <- read_csv("data/eqi_lbw_clean_df.csv")
```

#Prelim model just with EQI as exposure and low bw as outcome ##Poisson model

```
######Fit a Poisson model (not accounting for overdispersion)
mod1_p = glm(num_low_birthweight_births ~ eqi,
          data=eqi_lbw_clean_df,
          family=poisson,
          offset=log_live_births)

summary(mod1_p)
```

```
##
## Call:
## glm(formula = num_low_birthweight_births ~ eqi, family = poisson,
##     data = eqi_lbw_clean_df, offset = log_live_births)
```

```
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.4911216  0.0007558 -3295.84   <2e-16 ***
## eqi         -0.0334156  0.0008549   -39.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 77658  on 2994  degrees of freedom
## Residual deviance: 76134  on 2993  degrees of freedom
## AIC: 97828
##
## Number of Fisher Scoring iterations: 4
```

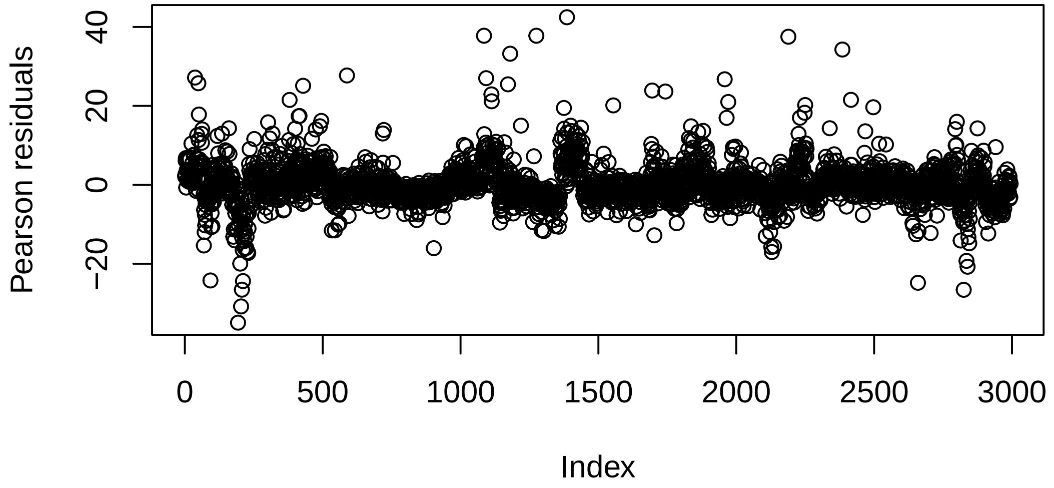
```
sum(resid(mod1_p,type="pearson")^2)/mod1_p$df.residual
```

```
## [1] 26.26938
```

```
#yes dispersion is potential problem bc scale > 1

#goodness of fit
pchisq(mod1_p$deviance, mod1_p$df.residual, lower.tail=F)
```

```
## [1] 0
```

```
#seeing a lack of fit for Poisson model

#Pearson residual plot
plot(resid(mod1_p,type="pearson"),ylab="Pearson residuals")
```



```
#also potential issue with outliers
```

#Prelim model just with EQI as exposure and low bw as outcome ##Quasipoisson model

```
######Fit a Quasipoisson model (which acounts for overdispersion)
#fit the Poisson model (accounting for overdispersion)
#no offset bc exposure unit is already same here
mod1_qp = glm(num_low_birthweight_births ~ eqi,
          data=eqi_lbw_clean_df,
          family=quasipoisson,
          offset=log_live_births)

summary(mod1_qp)
```

```
##
## Call:
## glm(formula = num_low_birthweight_births ~ eqi, family = quasipoisson,
##     data = eqi_lbw_clean_df, offset = log_live_births)
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.491122   0.003874 -643.046  < 2e-16 ***
## eqi         -0.033416   0.004382   -7.626 3.23e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 26.26938)
##
##     Null deviance: 77658  on 2994  degrees of freedom
## Residual deviance: 76134  on 2993  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

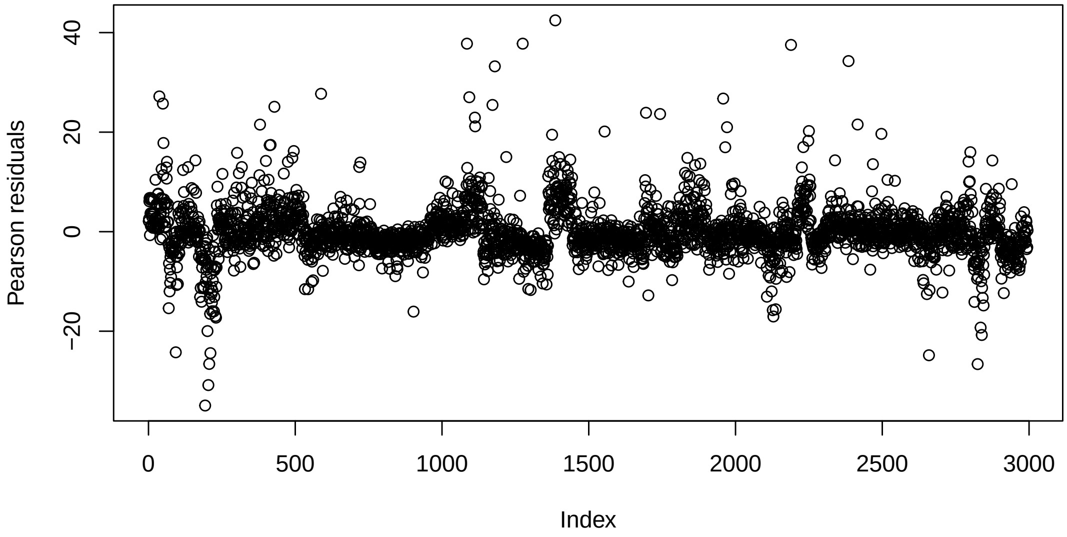
```
sum(resid(mod1_qp,type="pearson")^2)/mod1_qp$df.residual
```

```
## [1] 26.26938
```

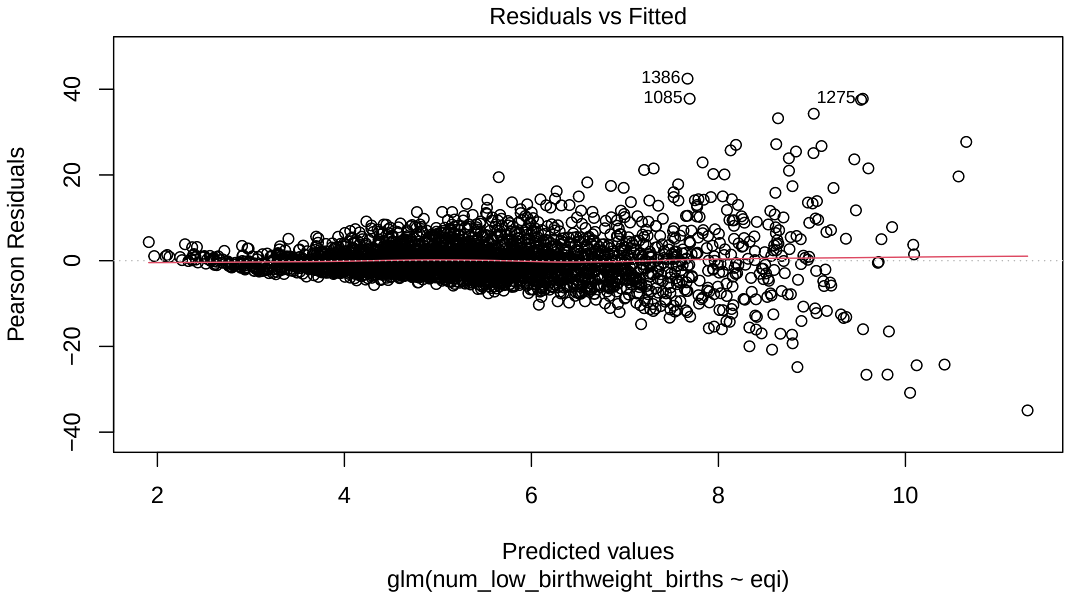
```
#yes dispersion

#goodness of fit
pchisq(mod1_qp$deviance, mod1_qp$df.residual, lower.tail=F)
```

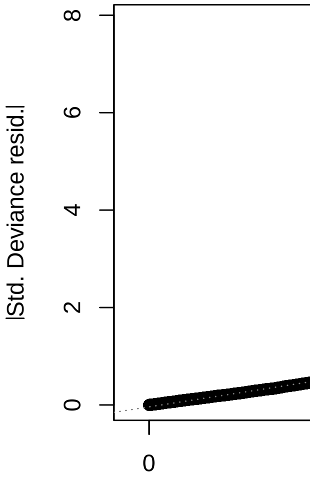
```
## [1] 0
```

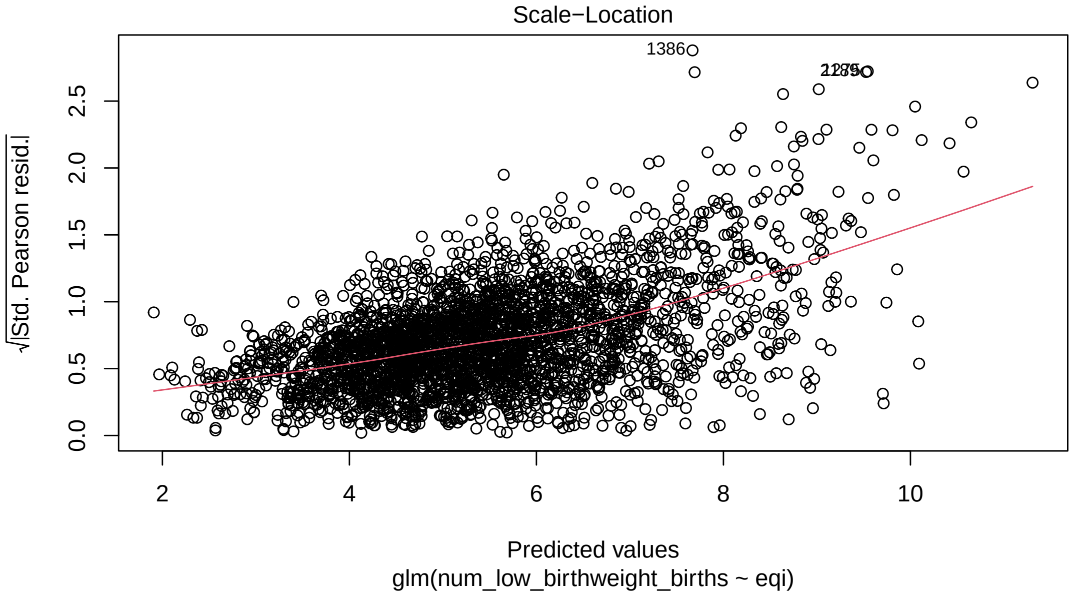
```
#seeing a lack of fit with the quasipoisson

#Pearson residual plot
plot(resid(mod1_qp,type="pearson"),ylab="Pearson residuals")
```

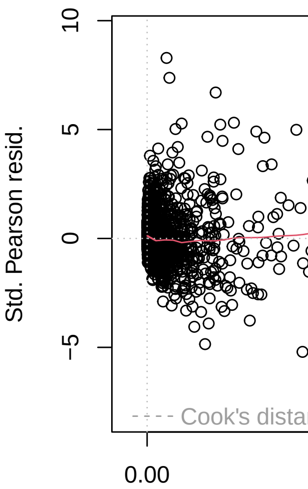


```
#also potential issue with outliers

plot(mod1_qp)
```









```
#looks perhaps nonlinear?
```

#Prelim model just with EQI as exposure and low bw as outcome ##Penalized Spline

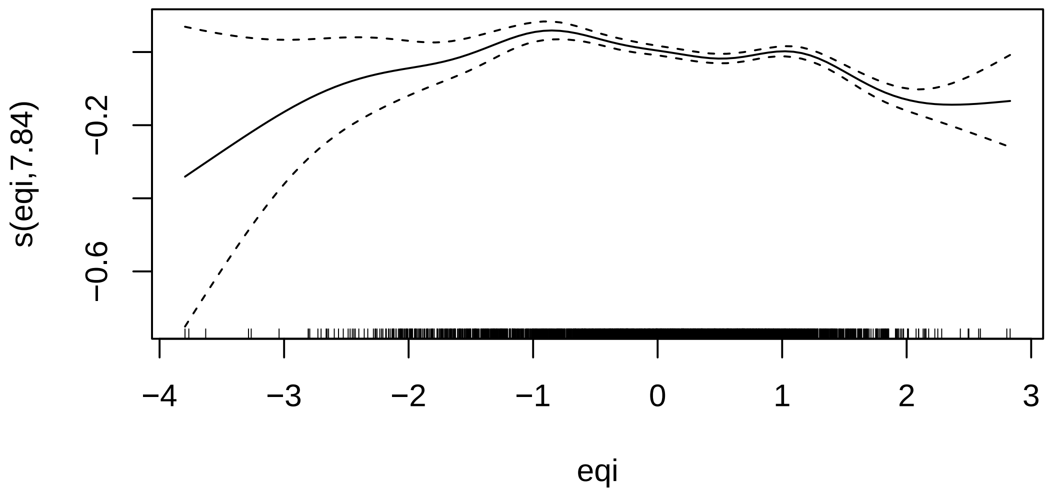
```
mod1_qp_nl <- gam(num_low_birthweight_births ~ s(eqi),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod1_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(eqi)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.499474   0.004047  -617.7   <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(eqi) 7.837  8.623 12.59  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.98   Deviance explained = 3.84%
## GCV = 25.082  Scale est. = 25.783    n = 2995
```

```r
#this is the penalty estimated by the model
mod1_qp_nl$sp
```

```
##   s(eqi)
## 22.53531
```

```r
plot(mod1_qp_nl)
```



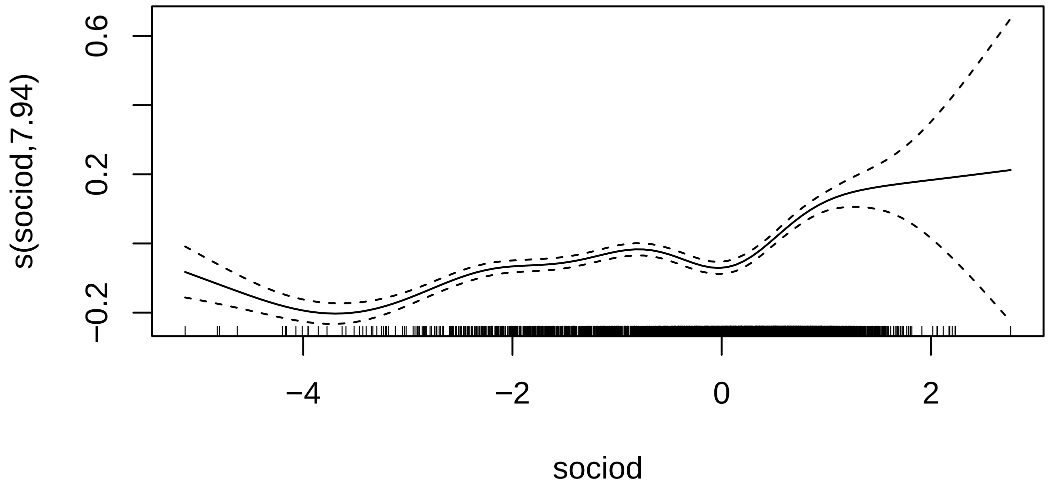
Seeing nonlinear relationship between eqi and num_low_birthweight_births

#Nonlinear checks of other indices and the low bw outcome; unadjusted

```r
##SOCIAL
mod2_qp_nl <- gam(num_low_birthweight_births ~ s(sociod),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod2_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(sociod)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.446520   0.005652  -432.8   <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(sociod) 7.94   8.57 38.58  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.984   Deviance explained = 10.7%
## GCV = 23.287  Scale est. = 24.963    n = 2995
```

```
plot(mod2_qp_nl)
```

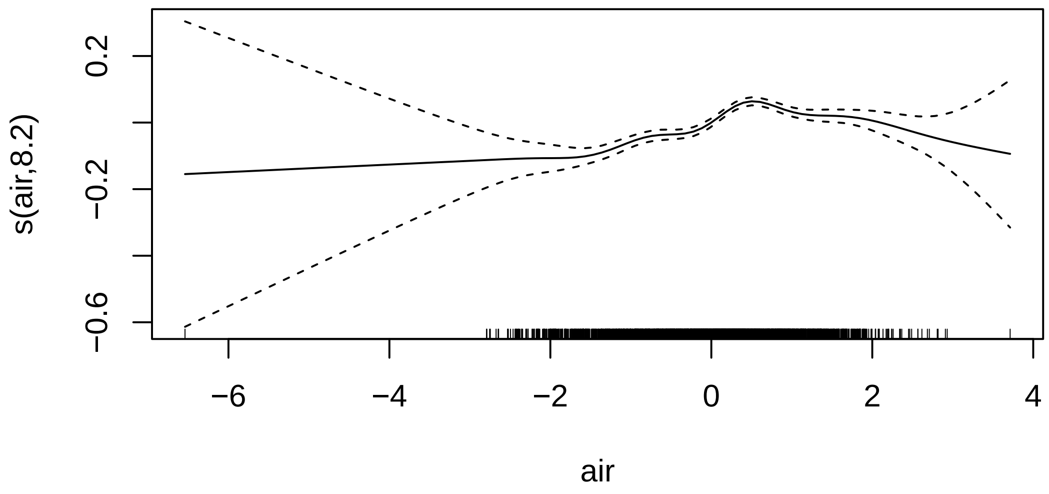


```
#nonlinear

##AIR
mod3_qp_nl <- gam(num_low_birthweight_births ~ s(air),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod3_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(air)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.509541   0.003389  -740.5   <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(air) 8.195  8.835 24.23  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.984   Deviance explained = 6.96%
## GCV = 24.274  Scale est. = 24.752    n = 2995
```

```
plot(mod3_qp_nl)
```

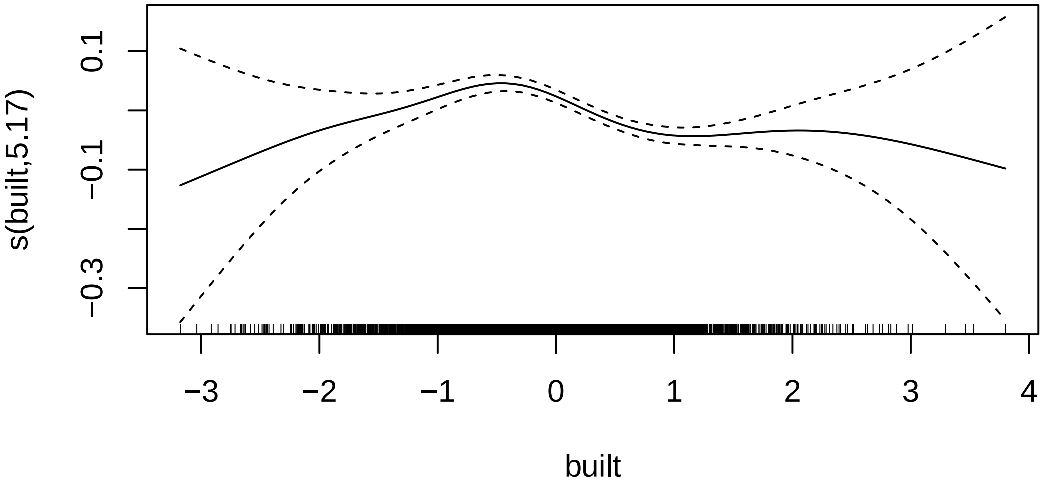


```
#nonlinear

##BUILT
mod4_qp_nl <- gam(num_low_birthweight_births ~ s(built),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod4_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(built)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.504570   0.004227  -592.5   <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(built) 5.166  6.331 14.43  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.979   Deviance explained = 3.18%
## GCV = 25.208  Scale est. = 25.702    n = 2995
```

```
plot(mod4_qp_nl)
```

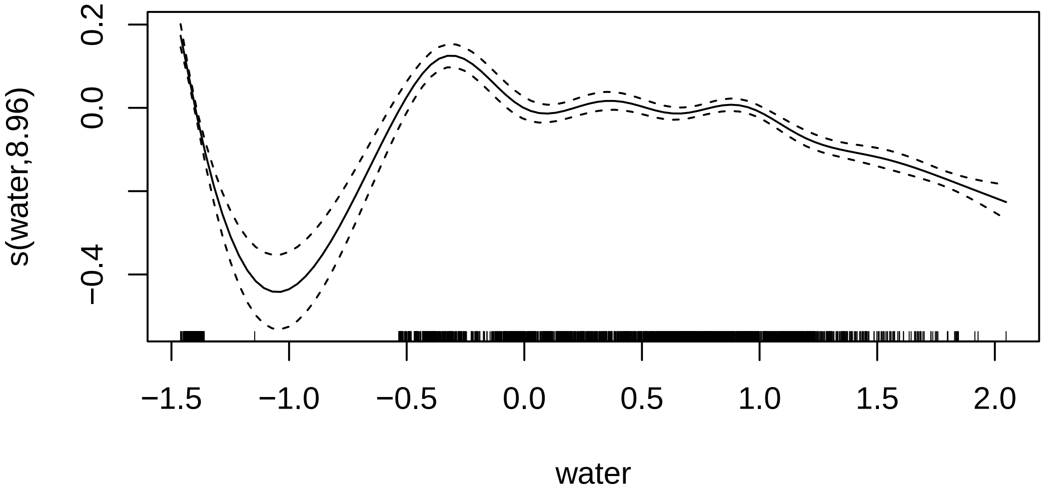


```
#nonlinear

##Water
mod5_qp_nl <- gam(num_low_birthweight_births ~ s(water),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod5_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(water)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.484860   0.003393  -732.4   <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(water) 8.963      9 70.08  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.986   Deviance explained = 18.3%
## GCV = 21.332  Scale est. = 22.134    n = 2995
```

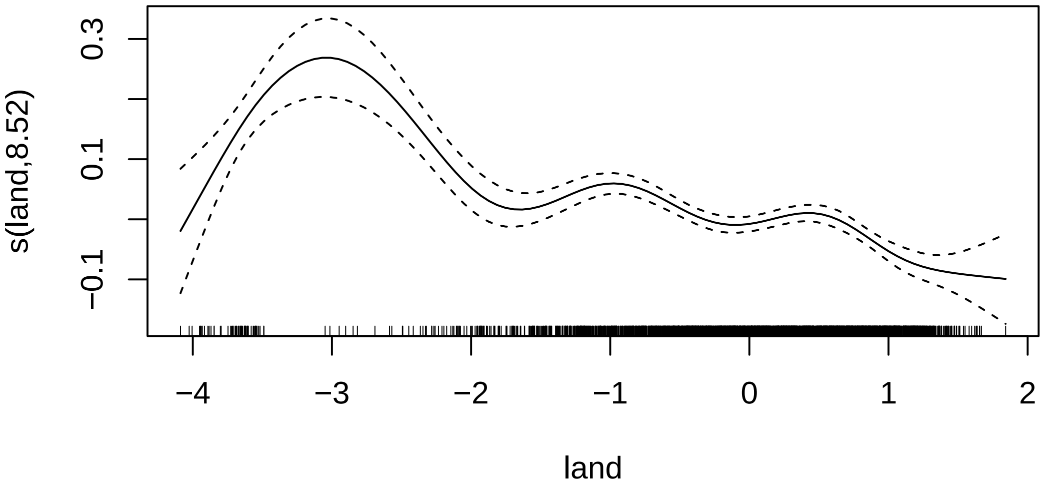
```
plot(mod5_qp_nl)
```



```
#nonlinear
#water index was weird distribution; establish a cutoff?

##Land
mod6_qp_nl <- gam(num_low_birthweight_births ~ s(land),
                  family = "quasipoisson",
                  offset=log_live_births,
                  data = eqi_lbw_clean_df)

summary(mod6_qp_nl)
```

```
##
## Family: quasipoisson
## Link function: log
##
## Formula:
## num_low_birthweight_births ~ s(land)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.516455   0.003486  -721.9   <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(land) 8.519  8.932 22.24  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.98   Deviance explained = 6.43%
## GCV = 24.418  Scale est. = 24.98     n = 2995
```

```
plot(mod6_qp_nl)
```



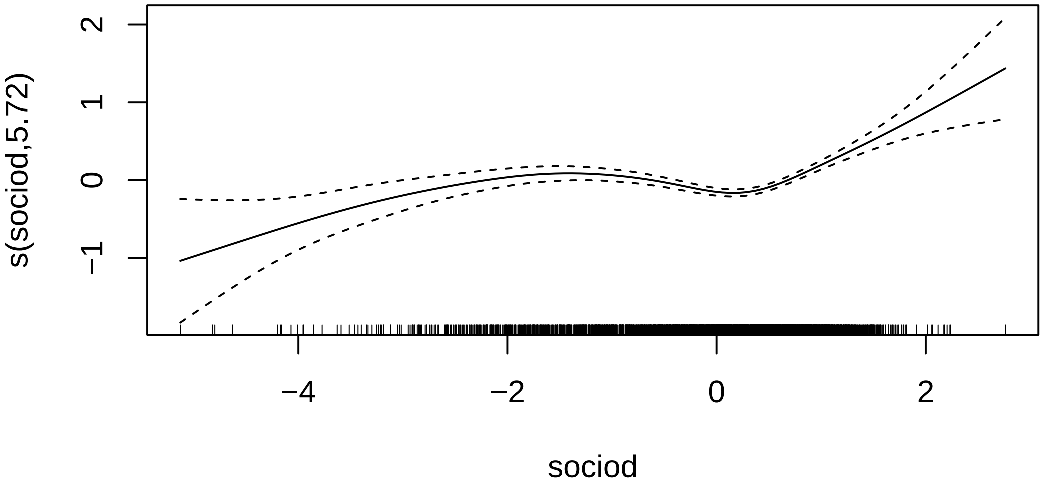
```
#nonlinear
```

#Are there nonlinear relationships between the 5 subdomain indices? ##penalized spline #using built index as the outcome for now

```
##SOCIAL
mod7_qp_nl <- gam(built ~ s(sociod),
                  data = eqi_lbw_clean_df)

summary(mod7_qp_nl)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## built ~ s(sociod)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04333    0.01672   2.592   0.0096 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##            edf Ref.df     F p-value
## s(sociod) 5.72  6.945 16.81  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0381   Deviance explained = 3.99%
## GCV = 0.83928  Scale est. = 0.83739   n = 2995
```

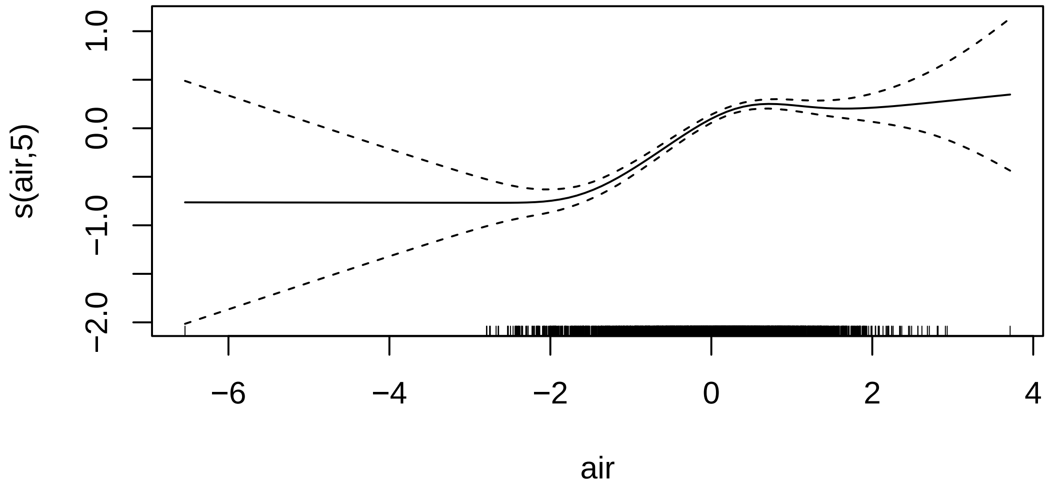
```
plot(mod7_qp_nl)
```



```
#nonlinear

##AIR
mod8_qp_nl <- gam(built ~ s(air),
                  data = eqi_lbw_clean_df)

summary(mod8_qp_nl)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## built ~ s(air)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04333    0.01611    2.69  0.00718 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##          edf Ref.df     F p-value
## s(air) 5.005  6.185 57.94  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.107   Deviance explained = 10.9%
## GCV = 0.77872  Scale est. = 0.77716   n = 2995
```

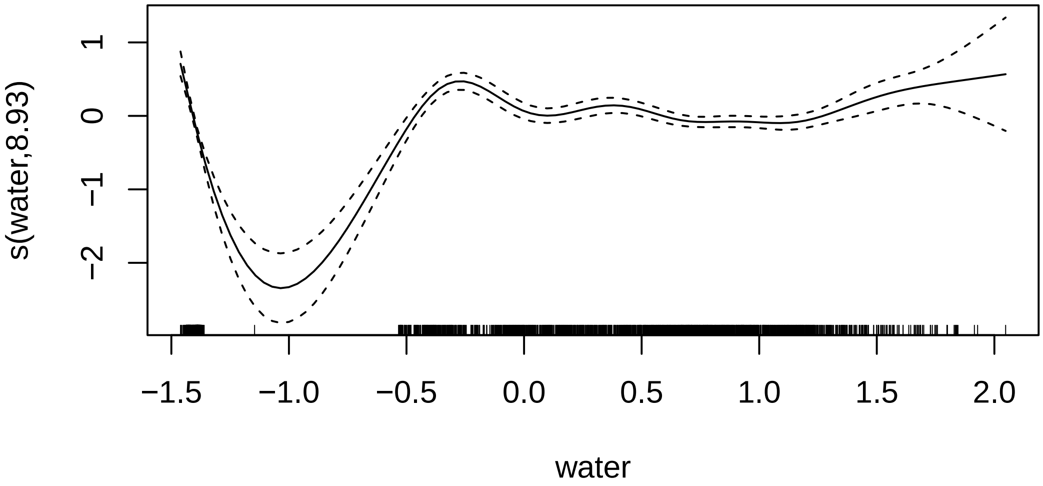
```
plot(mod8_qp_nl)
```



```
#nonlinear

##Water
mod9_qp_nl <- gam(built ~ s(water),
                  data = eqi_lbw_clean_df)

summary(mod9_qp_nl)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## built ~ s(water)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04333    0.01667   2.599   0.0094 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df     F p-value
## s(water) 8.933  8.999 16.06  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0434   Deviance explained = 4.63%
## GCV = 0.8355  Scale est. = 0.83273   n = 2995
```

```
plot(mod9_qp_nl)
```

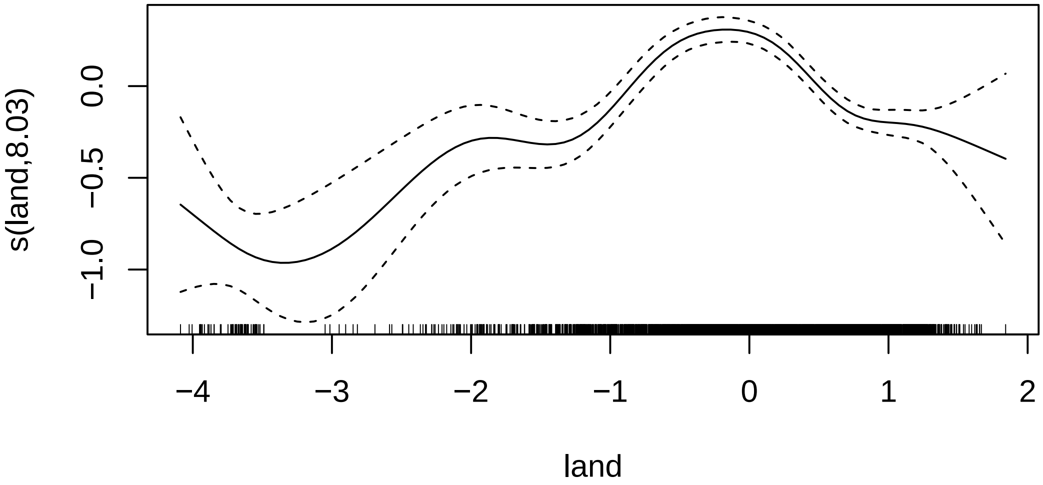


```
#nonlinear; again water has weird distribution; use cutoff?

##Land
mod10_qp_nl <- gam(built ~ s(land),
                   data = eqi_lbw_clean_df)

summary(mod10_qp_nl)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## built ~ s(land)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.04333    0.01642   2.639  0.00835 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Approximate significance of smooth terms:
##           edf Ref.df  F p-value
## s(land) 8.028  8.756 27  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0725   Deviance explained =  7.5%
## GCV = 0.80985  Scale est. = 0.8074    n = 2995
```

```
plot(mod10_qp_nl)
```



```
#very nonlinear
```

"'