finalAnalysis_v0.0.1

Brady M. Chisholm

2024-11-25

This is preliminary analysis, v0.0.1 and will be updated further.

A note about the data: Each pupil dilation measurement for a subject from that specific combination of conditions (4 SNR condition, 5 quintiles), averaged from 7-9 seconds. This generates 20 measurements per subject. For example, the first subject in this data is P02, and their data occupies the first 20 rows. P03 follows, etc.

Ideally this analysis is not done with an average of conditions for each, and is thousands of rows longer with trial-by-trial data. This is the first deep dive into this data, and more precise statistics can be done later as an extension on this work.

First load in data

```
pupilDat <- read.csv("~/Desktop/FA24 Syllabi:Schedules/PSY4802_Using_R_to_Create_Reproducible_Research/Final Project/code/quintileData.csv")

#FIXME github dat and use permalink instead
```

Check spread of data, and basic summary statistics

```
library(dplyr) # in ggplot?
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(mgcv) # best documentation available for GAMs that I found.
```

```
## Loading required package: nlme
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
## This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.
```

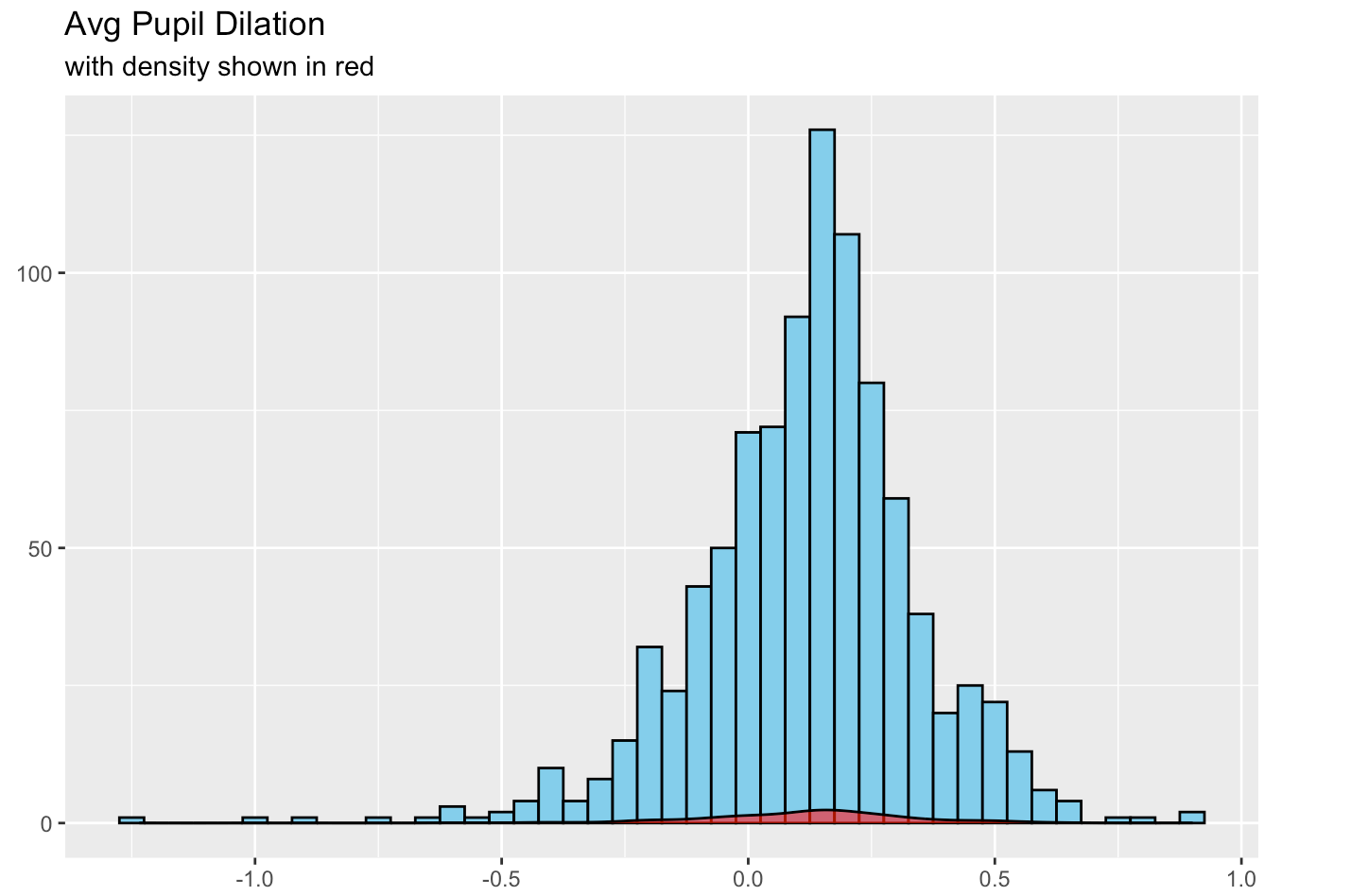
```
# summary stats table
print(as.data.frame(summary_stats <- pupilDat %>%
  group_by(SNR, Quintile) %>%
  summarise(
    Mean_Dilation = mean(AvgPupilDilation, na.rm = TRUE),
    Med_Dilation = median(AvgPupilDilation, na.rm = TRUE),
    SD_Dilation = sd(AvgPupilDilation, na.rm = TRUE),
    Min_Dilation = min(AvgPupilDilation, na.rm = TRUE),
    Max_Dilation = max(AvgPupilDilation, na.rm = TRUE),
    Count = n()),n=Inf))
```

```
## `summarise()` has grouped output by 'SNR'. You can override using the `.groups`
## argument.
```

```
##    SNR Quintile Mean_Dilation Med_Dilation SD_Dilation Min_Dilation
## 1  -10        1    0.21654252   0.20556896   0.1890342   -0.2130167
## 2  -10        2    0.15022221   0.14111626   0.2367548   -0.5047985
## 3  -10        3    0.23316331   0.18979226   0.2087505   -0.1426734
## 4  -10        4    0.14461768   0.17685099   0.2150664   -0.6395494
## 5  -10        5    0.16411442   0.16289793   0.2283283   -0.7329934
## 6   -5        1    0.15477986   0.15399755   0.1883877   -0.1674221
## 7   -5        2    0.11965716   0.13764337   0.1902422   -0.2361186
## 8   -5        3    0.10712631   0.11684919   0.2171754   -0.5018844
## 9   -5        4    0.08799807   0.14190823   0.2324006   -0.6106649
## 10  -5        5    0.16827295   0.17267720   0.2592883   -0.5865744
## 11   0        1    0.11397151   0.11907812   0.2115271   -0.4285123
## 12   0        2    0.07391466   0.13399513   0.2460966   -0.5657142
## 13   0        3    0.10259094   0.13780567   0.1834024   -0.4239136
## 14   0        4    0.06347232   0.07692387   0.2157571   -0.4470125
## 15   0        5    0.11758198   0.15538815   0.2251492   -0.8878726
## 16   5        1    0.11868438   0.14527973   0.2109090   -0.4288403
## 17   5        2    0.10241984   0.09956788   0.2077779   -0.4085787
## 18   5        3    0.03982967   0.10560630   0.2740865   -1.2347902
## 19   5        4    0.04328208   0.10640333   0.2533405   -1.0138579
## 20   5        5    0.08108403   0.06764727   0.1895675   -0.5892080
##    Max_Dilation Count
## 1     0.7392612    47
## 2     0.6628847    47
## 3     0.8119117    47
## 4     0.5049810    47
## 5     0.6301731    47
## 6     0.6182271    47
## 7     0.5639192    47
## 8     0.5870940    47
## 9     0.4339413    47
## 10    0.9012528    47
## 11    0.5465806    47
## 12    0.6074791    47
## 13    0.4975189    47
## 14    0.5809652    47
## 15    0.4805840    47
## 16    0.5057905    47
## 17    0.5648976    47
## 18    0.4750412    47
## 19    0.4431143    47
## 20    0.4708757    47
```

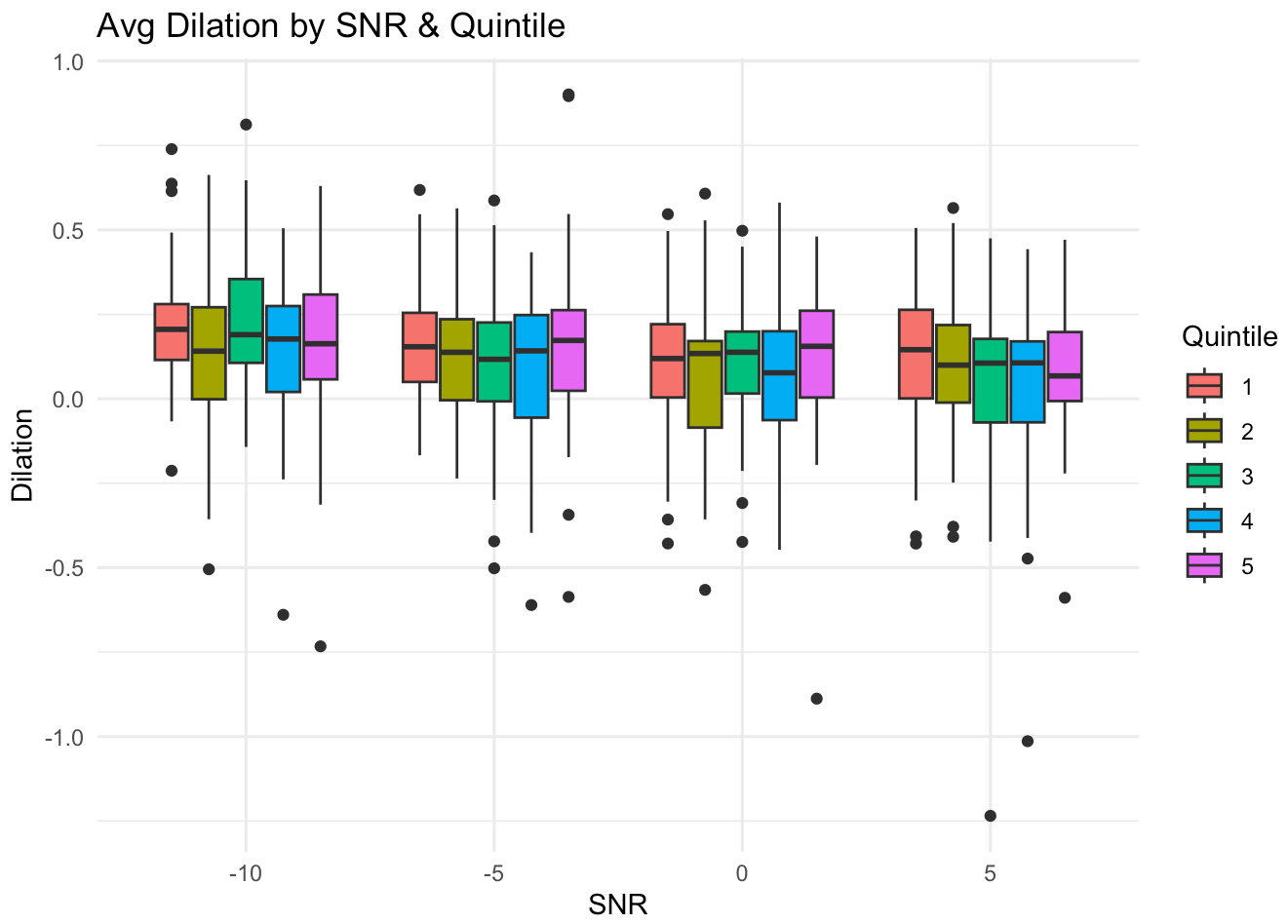
```
# tibble gets truncated for some reason...
```

```
# Check distribution of pupilDilation w/ histogram
ggplot(pupilDat, aes(x = AvgPupilDilation)) +
  geom_histogram(binwidth = 0.05, color = "black", fill = "skyblue") +
  geom_density(alpha = 0.6, fill = "red") +
  labs(title = "Avg Pupil Dilation", x = NULL, y = NULL,
       subtitle = "with density shown in red")
```



The histogram looks relatively well distributed, with maybe a small left-skew. This should not impact the GAM, but it is something to remember if we get any suspicious results

```
# plot spread of predictor vars against dilation
ggplot(pupilDat, aes(x = factor(SNR), y = AvgPupilDilation, fill = factor(Quintile))) +
  geom_boxplot() +
  labs(
      title = "Avg Dilation by SNR & Quintile",
       x = "SNR",
       y = "Dilation",
       fill = "Quintile") +
  theme_minimal()
```




```
# what trends or relationships exist in the data?
ggplot(pupilDat, aes(x = SNR, y = AvgPupilDilation)) +
  geom_point(alpha = 0.5) +
  geom_smooth(color = "blue", se = TRUE) + # defaulted to loess? Why?
  labs(title = "Avg Pupil Dilation vs. SNR", x = "SNR Level", y = "Avg Pupil Dilation")
+
  theme_minimal()
```

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

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : pseudoinverse used at -10.075
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : neighborhood radius 10.075
```

```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : reciprocal condition number 4.7738e-15
```

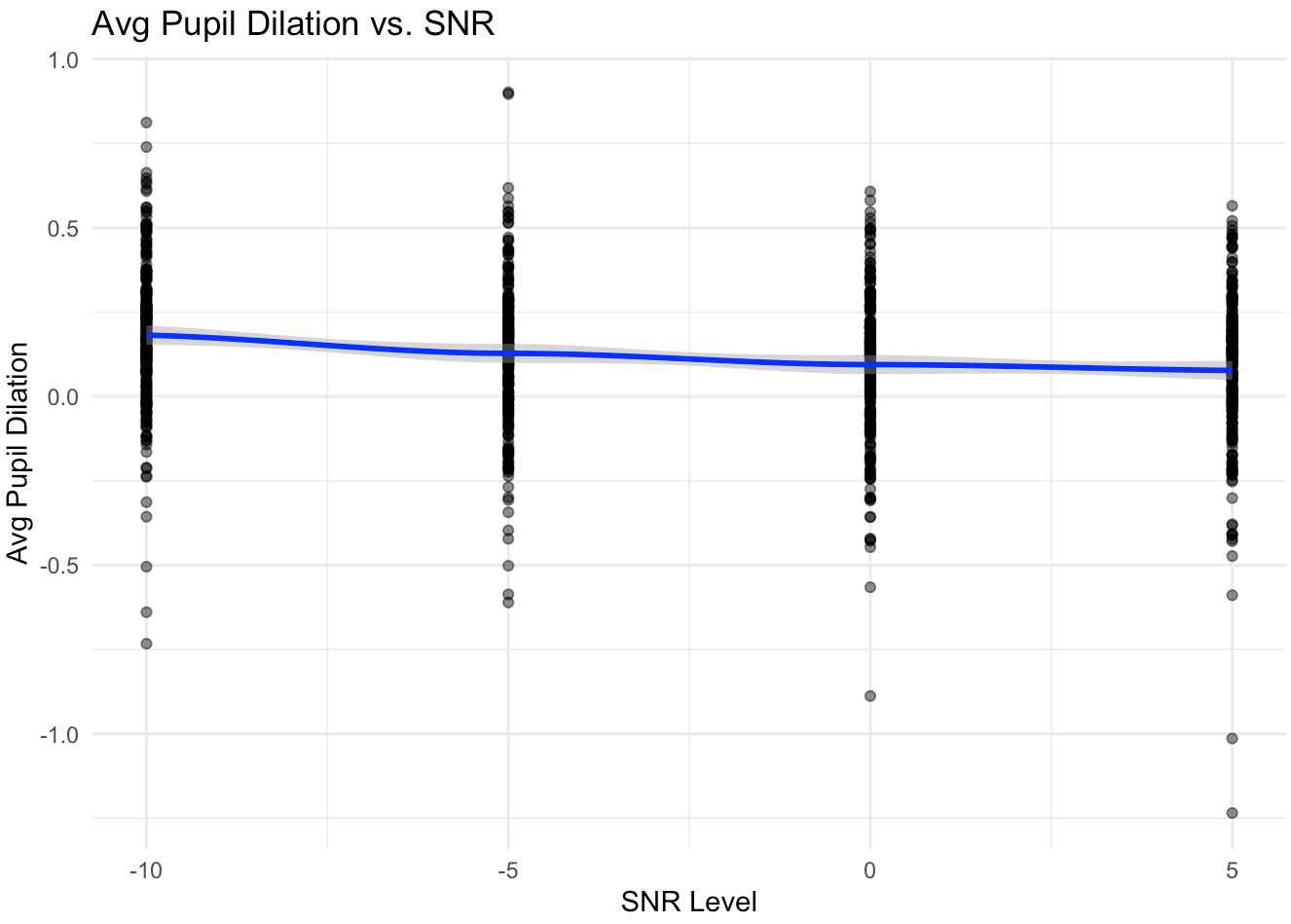
```
## Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
## : There are other near singularities as well. 101.51
```

```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used at
## -10.075
```

```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood radius
## 10.075
```

```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal condition
## number 4.7738e-15
```

```
## Warning in predLoess(object$y, object$x, newx = if (is.null(newdata)) object$x
## else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other near
## singularities as well. 101.51
```

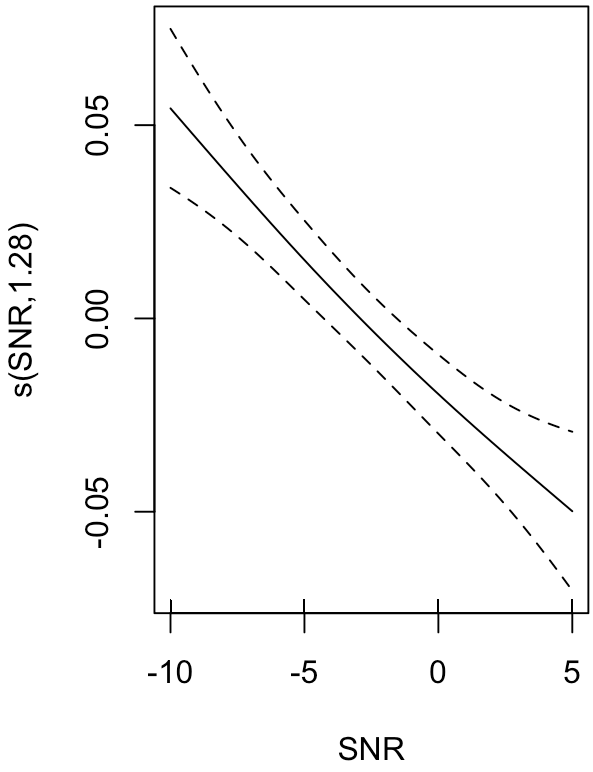


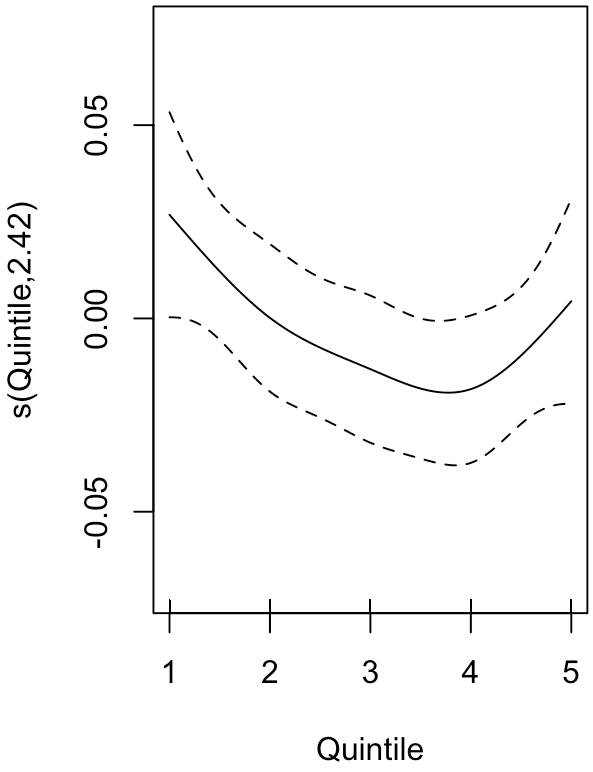
```
# REMEMBER k = # of knots. SNR has 4 levels, so 4 knots

# 1st model includes no interactions terms
gam_mod1 <- gam(AvgPupilDilation ~ s(SNR, k = 4) + s(Quintile, k = 5), data = pupilDat)
summary(gam_mod1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## AvgPupilDilation ~ s(SNR, k = 4) + s(Quintile, k = 5)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12017    0.00718   16.74   <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(SNR)      1.279  1.498 18.039 1.69e-06 ***
## s(Quintile) 2.421  2.922  1.958    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0342   Deviance explained =  3.8%
## GCV = 0.0487  Scale est. = 0.048456  n = 940
```

```
plot(gam_mod1, pages = 1)
```

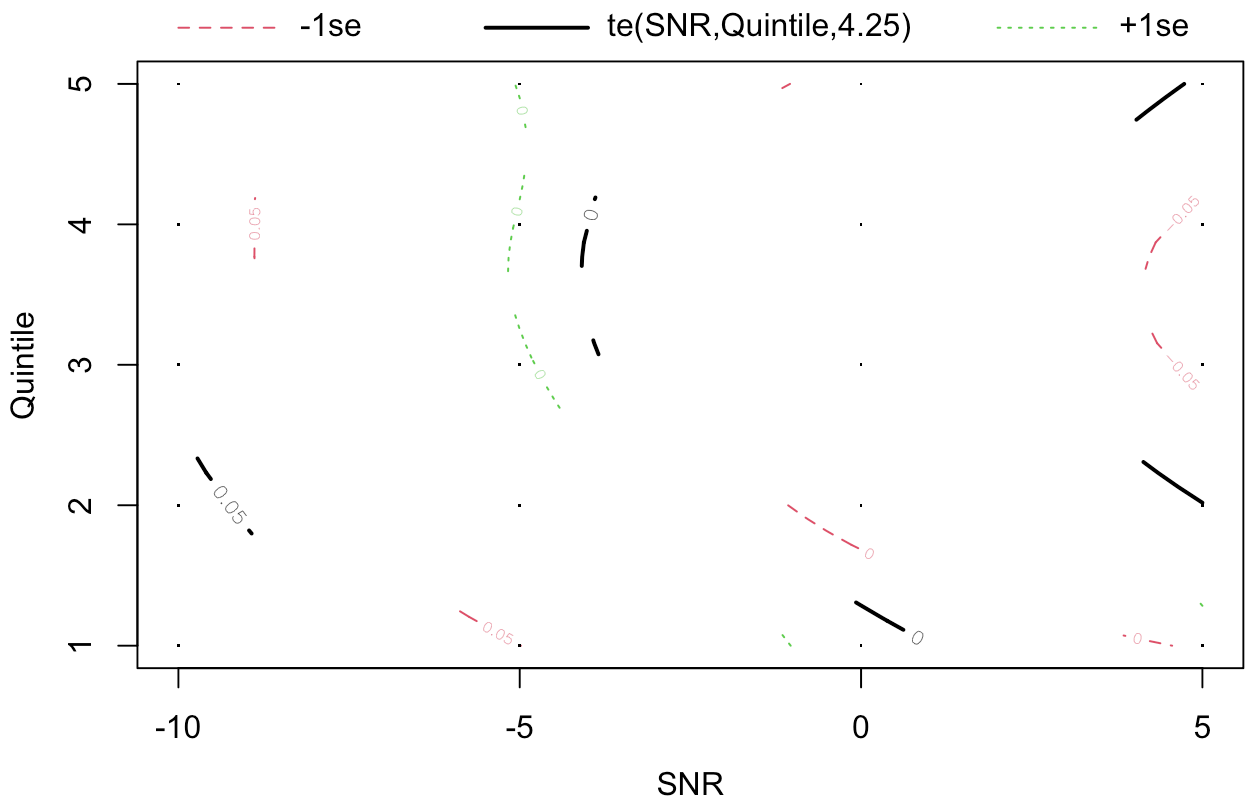





```
# interaction between SNR & quintile
gam_mod2 <- gam(AvgPupilDilation ~ te(SNR, Quintile,k = 4), data = pupilDat)
summary(gam_mod2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## AvgPupilDilation ~ te(SNR, Quintile, k = 4)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12017    0.00719   16.71   <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
## te(SNR,Quintile) 4.252  4.943 6.645 5.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0315   Deviance explained = 3.59%
## GCV = 0.048862  Scale est. = 0.048589  n = 940
```

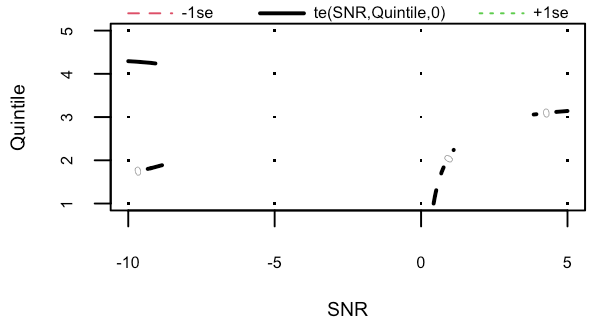
```
plot(gam_mod2, pages = 1)
```

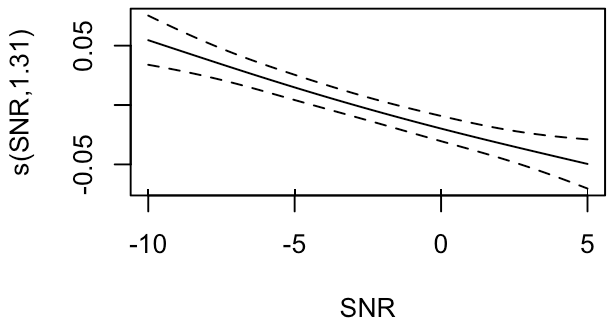


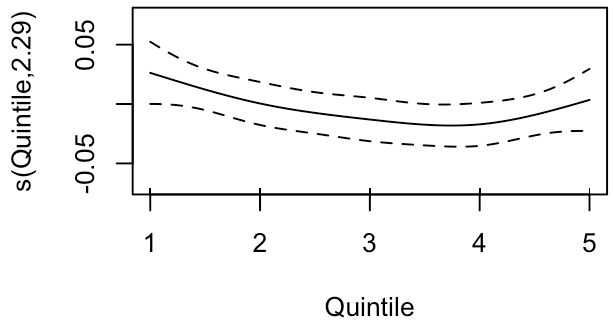
```
gam_mod3 <- gam(AvgPupilDilation ~ te(SNR, Quintile, k = 4) + s(SNR, k=4) + s(Quintile, k = 5), data = pupilDat)
summary(gam_mod3)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## AvgPupilDilation ~ te(SNR, Quintile, k = 4) + s(SNR, k = 4) +
##     s(Quintile, k = 5)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.12017    0.00718   16.74   <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
## te(SNR,Quintile) 6.449e-08  9.000  0.000    0.546
## s(SNR)           1.313e+00  1.551 17.371 1.93e-06 ***
## s(Quintile)      2.288e+00  2.775  2.032    0.138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.0341   Deviance explained = 3.78%
## GCV = 0.0487  Scale est. = 0.048461  n = 940
```

```
plot(gam_mod3, pages = 1)
```







Compare some of the initial models we've made

```
AIC(gam_mod1,gam_mod2,gam_mod3)
```

```
##                df       AIC
## gam_mod1 5.700388 -171.1715
## gam_mod2 6.251989 -168.0588
## gam_mod3 5.601514 -171.1745
```

Based on AIC, mod1 and mod3 are nearly identical at predicting pupil dilation response.

```
BIC(gam_mod1,gam_mod2,gam_mod3)
```

```
##                df       BIC
## gam_mod1 5.700388 -143.5482
## gam_mod2 6.251989 -137.7624
## gam_mod3 5.601514 -144.0302
```

```
model_comparison <- data.frame(
  Model = c("Mod_1", "Mod_2", "Mod_3"),
  AIC = c(AIC(gam_mod1), AIC(gam_mod2), AIC(gam_mod3)),
  BIC = c(BIC(gam_mod1), BIC(gam_mod2), BIC(gam_mod3))
)

knitr::kable(model_comparison, digits = 3, caption = "Comparison of AIC & BIC")
```

Comparison of AIC & BIC

Model	AIC	BIC
Mod_1	-171.172	-143.548
Mod_2	-168.059	-137.762
Mod_3	-171.174	-144.030

BIC indicated model 3 is the best fit, but by a negligible difference. BIC of 2-10 is considered significant enough for choosing a different model

We get significant results for all models when looking at the main effect of SNR on pupil dilation.

Quintile is not significant in any of these models, however it is likely due to the fact that this data is 940 rows and should really be thousands with more raw data included over averages. To further this analysis, we would want to rework how we produce the data table from the project for statistics. This is done in MATLAB (as the raw data is in .mat format), and it beyond the scope of what I've done here, but is the next step for furthering this work.

Another step would be to bootstrap the data and see how results look when we have significantly more data.

In summary, initial results indicate that there is a