Project 1 PHP 2550 Final

2023-10-08

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

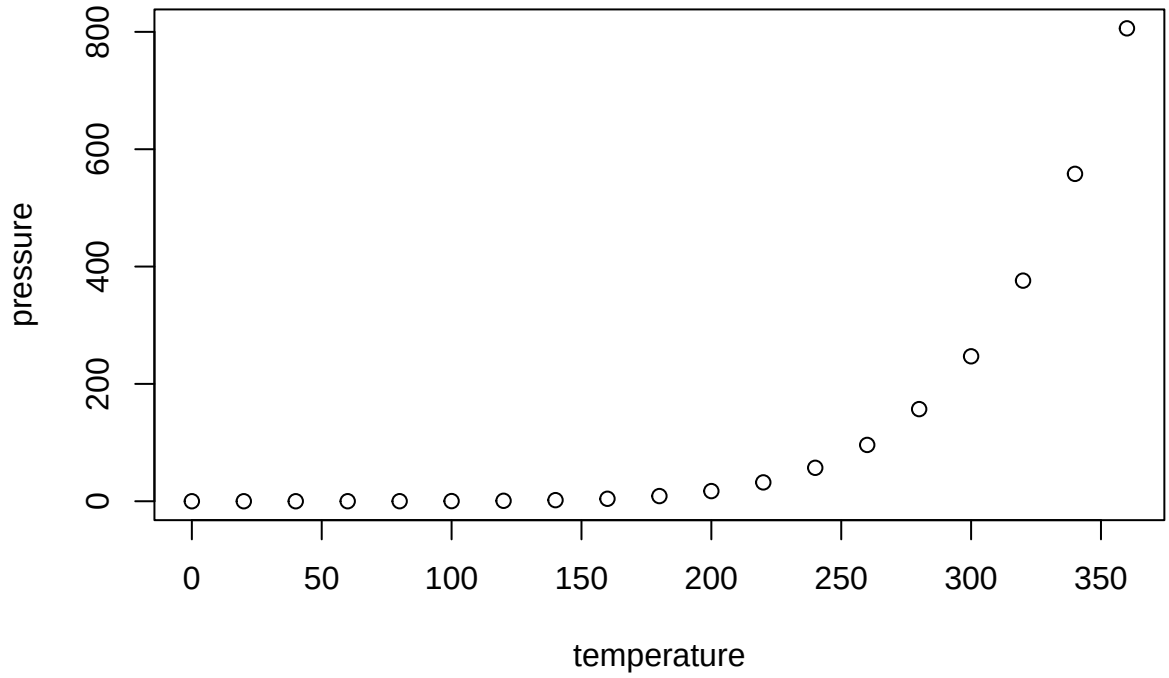
```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

```
# load the packages
library(tidyverse)
library(dplyr)
library(kableExtra)
library(gtsummary)
library(ggplot2)
library(naniar)
library(gridExtra)

# load the data
new_df<- read.csv("C:/Users/VGSCHOOL/project1.csv")
```

Including Plots

You can also embed plots, for example: Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.



There are 78 variables from 49 participants representing the demographics, socioeconomic status, education, and score measurements from survey of parent and child focusing on externalizing, self-regulation, and substance use. The variables also included smoking status during pre and postpartum (SDP) and follow up measurements on Environmental tobacco smoke (ETS) within a five year period after the child's birth.

First, I wanted to reformat the data to get rid of any outliers or missinginess that might create bias or skewed results. I first factored race, sex, education, income, and time to look at baseline characteristics. I found that there is one row with mostly NA values for measurements so I proceeded to delete this row. Then, I factored the time variables for smoking status at all time points. I continued checking the data for any outliers that may be included and they were multiple in the mom_numcig. I believed that there was an error in the data when it said a patient smokes 44,989 cigarettes in one day so I replaced the value with NA for missing so it wouldn't skew our results ,and changed two strings to integers "2 black and miles a day" and "none".I also added in NA values to columns in the SDP time variables and changed the NA values to 0 if the child answered that they had never consumed a substance before to account for it correctly in the missing data patterns in which will be conducted in the next step. In addition, I also changed the O values in swan_inattentive and swan_hyperactive to NA values as they were in the original dataset.

```
overall_missing<-miss_var_summary(new_df)
overall_missing
```

```
## # A tibble: 78 x 3
##    variable                  n_miss pct_miss
##    <chr>                      <int>    <dbl>
##  1 mom_smoke_pp1                 39     79.6
##  2 childasd                      28     57.1
##  3 mom_smoke_pp2                 20     40.8
##  4 pmq_parental_control          16     32.7
##  5 ppmq_parental_solicitation    15     30.6
```

```
##  6 swan_hyperactive               14     28.6
##  7 num_alc_30                     14     28.6
##  8 bpm_int                        14     28.6
##  9 pmq_parental_knowledge         14     28.6
## 10 pmq_parental_solicitation      14     28.6
## # i 68 more rows
```

```
 #Look at missing data for each parent
pct_na_r <- rowSums(is.na(new_df)) / ncol(new_df) * 100
row_na <- data.frame(parent_id = new_df$parent_id, pct_na = pct_na_r)
row_na <- row_na[row_na$pct_na > 70,]
print(row_na)
```

```
##    parent_id   pct_na
## 5      50502 74.35897
## 12     51202 71.79487
## 16     51602 76.92308
## 22     52302 71.79487
## 38     53902 73.07692
## 43     54402 76.92308
## 45     54602 76.92308
## 46     54702 73.07692
```

```
#Remove row with baseline values missing
row_to_remove <- 38
new_df <- new_df[-row_to_remove, ]


#Create imputation for smoking status for post-partum and during pregnancy
new_df$mom_smoke_32wk<- ifelse(new_df$mom_smoke_16wk == "1=Yes" &
                                 new_df$mom_smoke_22wk == "1=Yes",
                  "1=Yes", as.character(new_df$mom_smoke_32wk) )

new_df$mom_smoke_32wk<-factor(new_df$mom_smoke_32wk)

new_df$mom_smoke_pp1<- ifelse(new_df$mom_smoke_22wk == "1=Yes" &
                                new_df$mom_smoke_32wk == "1=Yes" &
                                new_df$mom_smoke_pp2 == "1=Yes",
                  "1=Yes", as.character(new_df$mom_smoke_pp1))
new_df$mom_smoke_pp1<-factor(new_df$mom_smoke_pp1)
```

Second, we would like to look at missing data patterns in baseline characteristics, SDP time variables, ETS time variables,score responses from survey, and across patients in general. There was an extraordinary amount of missing data all across the dataset but the highest variables of missingnesss were mom_smoke_pp1(79.59)% , childasd (57.14%), and mom_smoke_pp2(40.81%). The other smoking time variables for SDP and ETS all had missing percentages ranging from 22.45% to 14.29%. Furthermore,considering the cumulative nature of cigarette smoking and the nicotine addictive, I addressed this data gap by making the following assumptions. I assumed that if a parent's smoking behavior was known at 16 and 22 weeks, this behavior would remain consistent at 32 weeks (consistency of the addictive drug). Similarly, if we had information on a parent's smoking behavior at 22 and 32 weeks, as well as at postpartum visit 2, we assumed that this behavior would continue into postpartum visit 1. Although since there was a lot of missing data with the postpartum visits 1 & 2 and its not an accurate time point in which these visits occurred (such as 12weeks and 6 months) in the timeline so I have excluded them from my analysis when referring to post-partum smoking status and only focus on mom_smoke_pp12wk and mom_smoke_pp6mo which both have more complete data and time points for measurement.

Table 1: Table 1:Summary of SDP

SDP characteristics	#
mom_smoke_16wk	
1=Yes	12 (25%)
2=No	36 (75%)
mom_smoke_22wk	
1=Yes	13 (31%)
2=No	29 (69%)
Missing	6
mom_smoke_32wk	
1=Yes	11 (27%)
2=No	30 (73%)
Missing	7

The parental knowledge questions answered by the child had yielded high percentages of missingness around 28.5% to 26.5% and an assumption of lack of knowledge due to perspective of not being a parent/adult could account for these percentages. However, it's important to note that we refrained from making assumptions about environmental tobacco smoke (ETS) exposure variables, as they were subject to more frequent fluctuations compared to parental smoking behavior.

You can see across patients that there are 9 patients with missing data over 70%!!! Usually if we had a larger dataset we would probably remove all 9 patients but due to us having 49 patients I have decided to keep 8 out of 9 for the analysis. I have removed parent id 53902 because it has missingness in the baseline for SDP(mom_smoke_16wk). If a patient doesn't have baseline measurements it can cause a multitude of problems including bias, reduced statistical power, inaccurate estimates, and more. To avoid this you can use imputation or perform a sensitivity analysis but I believed it would just be best to do a complete case analysis. Below I have created tables showing the reformatted data and the baseline characteristics mean values, count, and missing values count.

```
table1_SDP<-new_df %>% select(mom_smoke_16wk,mom_smoke_22wk, mom_smoke_32wk) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 1:Summary of SDP",

col.names = c("SDP characteristics","#"))

table1_SDP
```

```
table2_PP<-new_df %>% select(mom_smoke_pp12wk,mom_smoke_pp6mo) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 2:Summary of Post-Partum",

col.names = c("Smoking PP characteristics","#"))
table2_PP
```

```
table_ETS<-new_df %>% select(smoke_exposure_12mo,smoke_exposure_12mo,smoke_exposure_2yr,smoke_exposure_3
                             smoke_exposure_4yr,smoke_exposure_5yr) %>%
tbl_summary(missing_text = "Missing") %>%
knitr::kable(caption = "Table 3:Summary of ETS",

col.names = c("Smoking Follow-Up characteristics","#"))
table_ETS
```

Table 2: Table 2:Summary of Post-Partum

Smoking PP characteristics	#
mom_smoke_pp12wk	
1=Yes	12 (29%)
2=No	30 (71%)
Missing	6
mom_smoke_pp6mo	
1=Yes	16 (40%)
2=No	24 (60%)
Missing	8

Table 3: Table 3:Summary of ETS

Smoking Follow-Up characteristics	#
smoke_exposure_12mo	
0	30 (77%)
1	9 (23%)
Missing	9
smoke_exposure_2yr	
0	28 (72%)
1	11 (28%)
Missing	9
smoke_exposure_3yr	
0	27 (71%)
1	11 (29%)
Missing	10
smoke_exposure_4yr	
0	28 (74%)
1	10 (26%)
Missing	10
smoke_exposure_5yr	
0	29 (74%)
1	10 (26%)
Missing	9

```r
#Find average composite score of follow up
new_df$avg_smoke_follow<-((rowSums(new_df[,c("smoke_exposure_6mo","smoke_exposure_12mo")] ==1,
                                   na.rm=T))/2
                          +(rowSums(new_df[,c("smoke_exposure_2yr","smoke_exposure_3yr",
                                              "smoke_exposure_4yr","smoke_exposure_5yr")] ==1,
                                    na.rm= T)))/5

new_df$cum_smoke_pp<-(rowSums(new_df[,c("mom_smoke_pp12wk","mom_smoke_pp6mo")]=="1=Yes",na.rm=T))

#Find cumulative score of SDP
new_df$cum_smoke_preg<-rowSums(new_df[,c("mom_smoke_16wk",
                               "mom_smoke_22wk",
                               "mom_smoke_32wk")] == "1=Yes", na.rm = T)
```

Next, I created composite scores for the SDP and post-partum smoking and average scores for follow up ETS smoking variables to simplify the data since they are measuring the same attribute of smoking status during these different points.

TIME TO ADDRESS THE AIMS: AIM 1: Examine effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing.

```r
#fit new model using cumulative count with alcohol use


model_cum_smoke_alc<- glm(alc_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
summary(model_cum_smoke_alc)
```

```
##
## Call:
## glm(formula = alc_ever ~ cum_smoke_preg, family = binomial(link = "logit"),
##     data = new_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8664  -0.4383  -0.4383  -0.4383   2.1866
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.2945     0.6563  -3.496 0.000472 ***
## cum_smoke_preg   0.5027     0.3534   1.422 0.154943
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 29.012  on 35  degrees of freedom
## Residual deviance: 27.081  on 34  degrees of freedom
##   (12 observations deleted due to missingness)
## AIC: 31.081
##
## Number of Fisher Scoring iterations: 5
```

```r
#create plot showing the probability predictions
new_cum_data <- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
```

```r
new_cum_data$Predicted_Probabilities<- predict(model_cum_smoke_alc, newdata = new_cum_data, type = "resp


# Create a plot for baseline and alcohol use predictions
alc_plot<-ggplot(data = new_cum_data, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child Alcohol Use",
    title = "Graph 1: Logistic Regression Predicted Probabilities"
  )
```

```r
#fit new model using cumulative count with marijuana use
model_cum_smoke_mj<- glm(mj_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
summary(model_cum_smoke_mj)
```

```
##
## Call:
## glm(formula = mj_ever ~ cum_smoke_preg, family = binomial(link = "logit"),
##     data = new_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7437  -0.2602  -0.2602  -0.2602   2.6088
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.3690     1.0431  -3.230  0.00124 **
## cum_smoke_preg   0.7417     0.4501   1.648  0.09937 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20.824  on 36  degrees of freedom
## Residual deviance: 17.917  on 35  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 21.917
##
## Number of Fisher Scoring iterations: 6
```

```r
#create plot showing the probability predictions
new_cum_data_mj<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_mj$Predicted_Probabilities<- predict(model_cum_smoke_mj, newdata = new_cum_data_mj, type =


# Create a plot for baseline and alcohol use predictions
mj_plot<-ggplot(data = new_cum_data_mj, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child Marijuana Use",
    title = "Graph 2: Logistic Regression Predicted Probabilities"
```

```
  )
```

```
#fit new model using cumulative count with ecig use
model_cum_smoke_ecig<- glm(e_cig_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
summary(model_cum_smoke_ecig)
```

```
##
## Call:
## glm(formula = e_cig_ever ~ cum_smoke_preg, family = binomial(link = "logit"),
##     data = new_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5834  -0.3452  -0.3452  -0.3452   2.3876
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.7906     0.8126  -3.434 0.000595 ***
## cum_smoke_preg   0.3687     0.4279   0.862 0.388950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20.824  on 36  degrees of freedom
## Residual deviance: 20.125  on 35  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 24.125
##
## Number of Fisher Scoring iterations: 5
```

```
#create plot showing the probability predictions
new_cum_data_ecig<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_ecig$Predicted_Probabilities<- predict(model_cum_smoke_ecig, newdata = new_cum_data_ecig,


# Create a plot for baseline and alcohol use predictions
ecig_plot<-ggplot(data = new_cum_data_ecig, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child E-cigarette Use",
    title = "Graph 3:Logistic Regression Predicted Probabilities"
  )
#fit new model using cumulative count with cigarette use

model_cum_smoke_cig<- glm(cig_ever ~ cum_smoke_preg, data = new_df, family = binomial(link = "logit"))
summary(model_cum_smoke_cig)
```

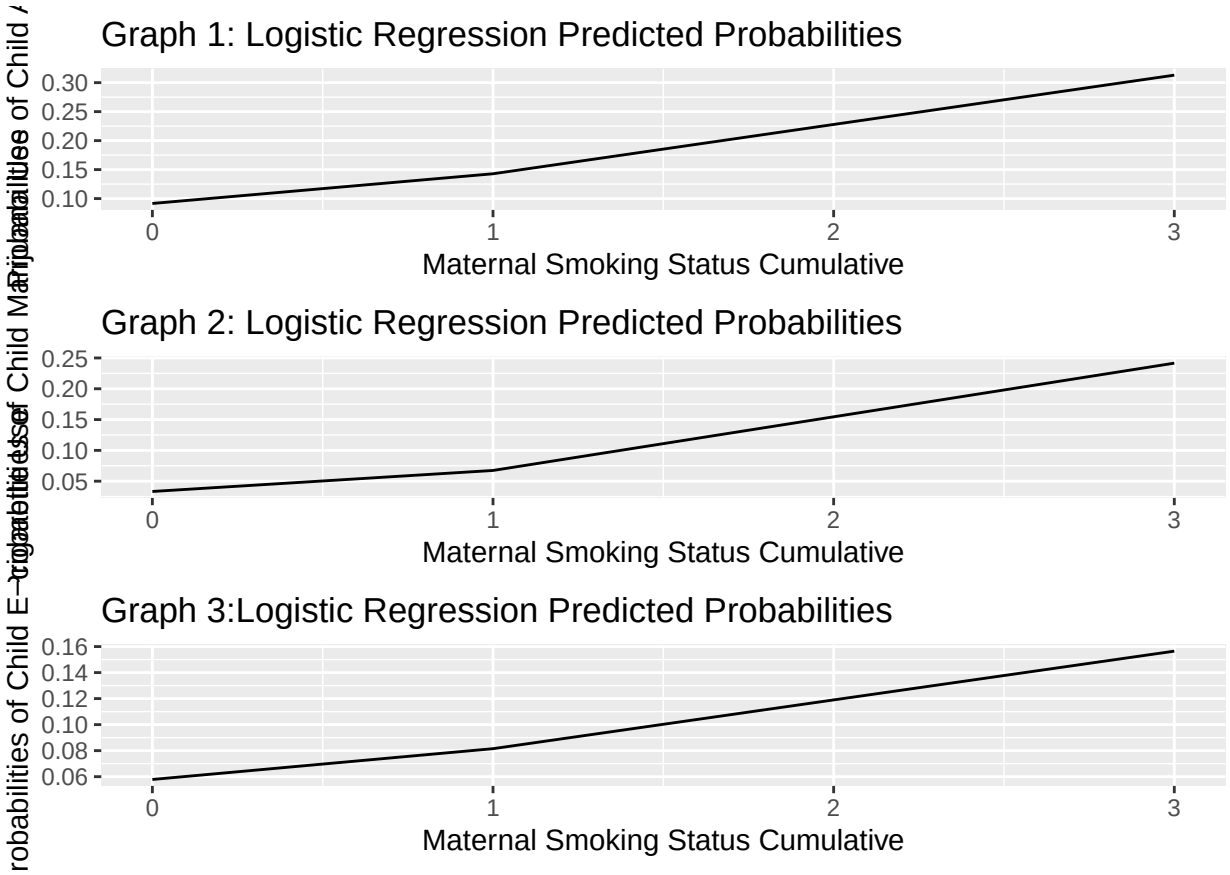
```
##
## Call:
## glm(formula = cig_ever ~ cum_smoke_preg, family = binomial(link = "logit"),
##     data = new_df)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51678  -0.00007   0.00000   0.00000   2.03933
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -28.637  10143.766  -0.003    0.998
## cum_smoke_preg     8.897   3381.255   0.003    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9.1946  on 36  degrees of freedom
## Residual deviance: 6.0283  on 35  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 10.028
##
## Number of Fisher Scoring iterations: 21
```

```r
#create plot showing the probability predictions
new_cum_data_cig<- data.frame(cum_smoke_preg = new_df$cum_smoke_preg)

# Predict probabilities using the model
new_cum_data_cig$Predicted_Probabilities<- predict(model_cum_smoke_cig, newdata = new_cum_data_cig, type


# Create a plot for baseline and alcohol use predictions
cig_plot<-ggplot(data = new_cum_data_cig, aes(x = cum_smoke_preg, y = Predicted_Probabilities)) +
  geom_line() +
  labs(
    x = "Maternal Smoking Status Cumulative",
    y = "Predicted Probabilities of Child cigarette Use",
    title = "Graph 4: Logistic Regression Predicted Probabilities"
)
grid.arrange(alc_plot, mj_plot, ecig_plot, heights = c(4,4,4))
```



First, I looked at the effects of SPD/ ETS on substance use in adolescents.I decided to look at the SDP and substance use effects using a GLM model because the presence of outliers or data that does not conform to normality assumptions. Also, GLMs allow for the modeling of nonlinear relationships between predictor variables and the response, making them versatile in capturing complex associations in the data. We see in the GLM models for SDP that for mj_ever,alc_ever, and ecig_ever that the baseline risk, was statistically significant (p = 0.00124,p=0.000472, p=0.000595), indicating that the likelihood of the three variables was significantly different from zero when no other predictor variables were considered. However, the cumulative score of SDP shows to possibly not be significant on substance use for children. The intercept was not significant for cig_ever.

So for next steps I think it would best and more appropriate to look at effects through summary tables instead of GLMs because the statistics can help you identify potential issues, such as skewness, outliers, or heteroscedasticity,to examine the assumptions underlying those models more before applying it. With making this next transition I evaluated the ETS follow-up on substance use on adolescents.

```
new_df %>% select(avg_smoke_follow, alc_ever, cig_ever, e_cig_ever,

mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_30) %>%

tbl_summary( by = avg_smoke_follow, missing_text = "Missing",
type = list (c(alc_ever, cig_ever, e_cig_ever,mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_30)
c(avg_smoke_follow) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(alc_ever, cig_ever, e_cig_ever,mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30,num_alc_3
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
```

Characteristic	**0**, N = 34	**0.1**, N = 2	**0.5**, N = 1	**0.6**, N = 1	**0.7**, N = 1	**0.8**, N =
alc_ever	0.13 (0.34)	0.50 (0.71)	NA (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	1	0	0	0
cig_ever	0.04 (0.21)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
e_cig_ever	0.09 (0.29)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
mj_ever	0.13 (0.34)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
num_cigs_30	0.00 (0.00)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
num_e_cigs_30	0.09 (0.42)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
num_mj_30	1.43 (4.42)	0.00 (0.00)	0.00 (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	11	0	0	0	0	0
num_alc_30	0.50 (2.13)	0.00 (0.00)	NA (NA)	0.00 (NA)	0.00 (NA)	0.00 (0.00)
Missing	12	0	1	0	0	0

This summary table displays the effects of the average follow up ETS on the substance use variables. We see that none of the variables alc-ever,cig_ever, e_cig_ever, mj_ever, num_cigs_30, num_e_cigs_30, num_mj_30, num_alc_30, for the children showed any statistical significance with the exposed tobacco smoke in their early years of childhood. We can see that there is a lot of missingness once again and that most people have an average score of 0 for the follow-up. There was a lot of variation in the the data

```
#stratified by cum SDP
tbl1_swan<- new_df[,c(32,33,81)] %>%
  tbl_summary(by = cum_smoke_preg) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing Swan factor score with SDP")
as_kable_extra(tbl1_swan)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 4: Compare Externalizing Swan factor score with SDP

Characteristic	**0**, N = 34	**1**, N = 3	**3**, N = 11	**p-value**
swan_hyperactive	5 (4, 11)	12 (9, 13)	15 (6, 19)	0.14
Unknown	10	0	3	
swan_inattentive	10 (7, 13)	10 (10, 11)	16 (12, 17)	0.2
Unknown	6	0	3	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
tbl1_erq<- new_df[,c(73,74,81)] %>%
  tbl_summary(by = cum_smoke_preg) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing ERQ factors score with SDP")
as_kable_extra(tbl1_erq)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 5: Compare Externalizing ERQ factors score with SDP

Characteristic	**0**, N = 34	**1**, N = 3	**3**, N = 11	**p-value**
erq_exp	2.50 (2.00, 3.00)	3.25 (3.25, 3.88)	2.63 (2.50, 3.56)	0.081
Unknown	9	0	3	
erq_cog	3.08 (2.58, 3.79)	3.83 (3.42, 4.00)	3.00 (3.00, 3.42)	0.5
Unknown	8	0	4	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
new_df %>% select(cum_smoke_preg, bpm_att, bpm_att_p, bpm_ext,

bpm_ext_p, bpm_int, bpm_int_p) %>%

tbl_summary( by = cum_smoke_preg, missing_text = "Missing",
type = list (c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ "continuous",
c(cum_smoke_preg) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ c(2, 2))) %>%
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
```

Characteristic	**0**, N = 34	**1**, N = 3	**3**, N = 11	**p-value**
bpm_att	2.50 (2.27)	2.33 (4.04)	4.88 (2.70)	0.084
Missing	8	0	3	
bpm_att_p	1.64 (1.87)	1.00 (1.00)	3.75 (2.82)	0.094
Missing	9	0	3	
bpm_ext	2.54 (1.79)	3.00 (3.61)	3.63 (2.13)	0.45
Missing	8	0	3	
bpm_ext_p	1.54 (2.55)	0.33 (0.58)	2.63 (2.62)	0.29
Missing	8	0	3	
bpm_int	2.54 (2.15)	4.00 (6.93)	2.75 (2.55)	0.82
Missing	10	0	3	
bpm_int_p	1.93 (2.40)	1.33 (1.15)	3.50 (2.88)	0.18
Missing	6	0	3	

First we look at the statistics from the swan inattentive and swan hyperactive variable. It shows the median score and its confidence interval For example, where the cumulative SDP is 1 for swan hyperactive we see the median is 12 with a narrow CI with a lower bound of 9 and upper bound of 13. We see that for the hyperactive variable that Cumulative SDP for 0 has a wider range between 4,1 with a median of 5 and cumulative score 3 has an even wider interval from 6 to 19 with a median of 15.The p-value for the comparison is 0.14, suggests that there is no statistically significant difference in the "swan_hyperactive" variable between the three SDP groups. For swan inattentive the p-value as well did not prove statistical significance with a p-value of 0.2 and the different medians and confidence intervals can be seen in the summary table above.

I then observed the Emotion Regulation Questionairre responses between the cumulative SDP and erq_exp did have a small p-value (0.081) but still proved not to be significant.Where the cumulative SDP is 1 for erq_exp we see the median is 3.25 with a narrow CI with a lower bound of the median value 3.25 and upper bound of 3.88. We see that for the erq_exp variable that Cumulative SDP for 0 has a wider range between 2,3 with a median of 2.5. The erq_cog had a higher p-value that proved not to be significant as well of 0.5.

For the Brief Problem Monitor responses none of the variables showed significance with the Cumulative SDP but I do want to highlight the small p-value variables which are bpm_att and bpm_att_p. Bpm_int had the most missingness between all the variables and in the last table you will be able to see each of the confidence

intervals broken down between the SDP cum scores and their confidence intervals.

```
#stratified by cum PP
tbl2_swan<- new_df[,c(32,33,80)] %>%
  tbl_summary(by = cum_smoke_pp) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing Swan factor score with PP")
as_kable_extra(tbl2_swan)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 6: Compare Externalizing Swan factor score with PP

Characteristic	**0**, N = 30	**1**, N = 8	**2**, N = 10	**p-value**
swan_inattentive	9 (7, 12)	15 (12, 17)	14 (12, 16)	0.083
Unknown	4	3	2	
swan_hyperactive	5 (4, 12)	11 (6, 17)	10 (6, 17)	0.3
Unknown	7	4	2	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
tbl2_erq<- new_df[,c(73,74,80)] %>%
  tbl_summary(by = cum_smoke_pp) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing ERQ factors score with PP")
as_kable_extra(tbl2_erq)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 7: Compare Externalizing ERQ factors score with PP

Characteristic	**0**, N = 30	**1**, N = 8	**2**, N = 10	**p-value**
erq_exp	2.50 (2.13, 3.13)	2.50 (2.00, 2.50)	3.13 (2.50, 3.75)	0.3
Unknown	7	3	2	
erq_cog	3.08 (2.75, 3.71)	4.33 (3.00, 5.00)	3.00 (3.00, 3.42)	0.3
Unknown	6	3	3	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
new_df %>% select(cum_smoke_pp, bpm_att, bpm_att_p, bpm_ext,

bpm_ext_p, bpm_int, bpm_int_p) %>%

tbl_summary( by = cum_smoke_pp, missing_text = "Missing",
type = list (c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ "continuous",
c(cum_smoke_pp) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ c(2, 2))) %>%
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
```

Characteristic	**0**, N = 30	**1**, N = 8	**2**, N = 10	**p-value**
bpm_att	2.42 (2.38)	2.80 (2.77)	4.88 (2.70)	0.087
Missing	6	3	2	
bpm_att_p	1.54 (1.93)	3.25 (3.20)	3.00 (2.27)	0.064
Missing	6	4	2	
bpm_ext	2.46 (1.91)	3.00 (1.87)	3.75 (2.31)	0.37
Missing	6	3	2	
bpm_ext_p	1.04 (2.29)	4.20 (1.92)	2.00 (2.56)	0.007
Missing	6	3	2	
bpm_int	2.30 (2.29)	4.00 (2.83)	3.25 (3.81)	0.37
Missing	7	4	2	
bpm_int_p	1.73 (2.44)	3.40 (1.95)	3.00 (2.73)	0.043
Missing	4	3	2	

First we look at the statistics from the swan inattentive and swan hyperactive variable. It shows the median score and its confidence interval but this time we'll be evaluating these effects from the cumulative score from post-partum smoking.For swan inattentive at cumulative score of 1 we see the median is 15 with a lower bound of 12 and upper bound of 17 We see that for the inattentive variable that Cumulative SPP for 0 has a range between 7and 12 with a median of 9 and cumulative score 2 has an even interval from 12 to 16 with a median of 14.The p-value is 0.083, suggests that there is no statistically significant difference in the "swan_inattentive" variable between the three SDP groups. For swan hyperactive the p-value as well did not prove statistical significance with a p-value of 0.3 and the different medians and confidence intervals can be seen in the summary table above.

I then observed the Emotion Regulation Questionairre responses between the cumulative SPP and erq_exp and exp_cog had the same p-value of 0.3 and which suggested no statistical significance.Where the cumulative SPP is 1 for erq_exp we see the median is 2.50 with a narrow CI with a lower bound of the median value 2.00 and upper bound of 2.50. We see that for the erq_cog variable that Cumulative SPP for 0 has a wider range between 2.75,3.71 with a median of 3.08.

For the Brief Problem Monitor responses only two variables showed statistical significance which was bpm_ext_p with a p-value of 0.007 and bpm_int_p with a p-value of 0.043. With the Cumulative SPP I do want to highlight other the small but non-significant variables which are bpm_att and bpm_att_p.

```
#stratified by average ETS
tbl3_swan<- new_df[,c(32,33,79)] %>%
  tbl_summary(by = avg_smoke_follow) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing Swan factor score with ETS")
as_kable_extra(tbl3_swan)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 8: Compare Externalizing Swan factor score with ETS

Characteristic	0, N = 34	0.1, N = 2	0.5, N = 1	0.6, N = 1	0.7, N = 1	0.8, N = 3	0.9, N = 1	1, N = 5	p-value
swan_hyperactive	5 (4, 12)	9 (6, 11)	20 (20, 20)	17 (17, 17)	18 (18, 18)	15 (14, 16)	1 (1, 1)	6 (2, 6)	0.067
Unknown	13	0	0	0	0	0	0	0	
swan_inattentive	9 (7, 13)	13 (10, 15)	15 (15, 15)	18 (18, 18)	17 (17, 17)	17 (15, 18)	4 (4, 4)	11 (10, 12)	0.2
Unknown	9	0	0	0	0	0	0	0	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
tbl3_erq<- new_df[,c(73,74,79)] %>%
  tbl_summary(by = avg_smoke_follow) %>%
  add_p() %>%
  sort_p() %>%
  modify_caption("Compare Externalizing ERQ factors score with ETS")
as_kable_extra(tbl3_erq)%>%
  kable_styling(full_width=T, latex_options = c('HOLD_position'))
```

Table 9: Compare Externalizing ERQ factors score with ETS

Characteristic	0, N = 34	0.1, N = 2	0.5, N = 1	0.6, N = 1	0.7, N = 1	0.8, N = 3	0.9, N = 1	1, N = 5	p-value
erq_exp	2.50 (2.00, 2.88)	2.75 (2.63, 2.88)	2.75 (2.75, 2.75)	NA (NA, NA)	2.50 (2.50, 2.50)	3.50 (3.00, 3.63)	3.75 (3.75, 3.75)	3.75 (3.50, 4.00)	0.13
Unknown	11	0	0	1	0	0	0	0	
erq_cog	3.17 (2.58, 4.00)	3.17 (3.08, 3.25)	3.00 (3.00, 3.00)	3.00 (3.00, 3.00)	3.00 (3.00, 3.00)	3.00 (2.75, 3.42)	3.83 (3.83, 3.83)	3.25 (3.00, 3.88)	>0.9
Unknown	11	0	0	0	0	0	0	1	

[1] Median (IQR)

[2] Kruskal-Wallis rank sum test

```
new_df %>% select(avg_smoke_follow, bpm_att, bpm_att_p, bpm_ext,

bpm_ext_p, bpm_int, bpm_int_p) %>%

tbl_summary( by = avg_smoke_follow, missing_text = "Missing",
type = list (c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ "continuous",
c(avg_smoke_follow) ~ "categorical"),
statistic = all_continuous() ~ "{mean} ({sd})",
digits = list(c(bpm_att, bpm_ext, bpm_int,
bpm_att_p, bpm_ext_p, bpm_int_p) ~ c(2, 2))) %>%
add_p(all_categorical () ~ "chisq.test", pvalue_fun = ~ style_pvalue(.x, digits = 2))
```

Characteristic	**0**, N = 34	**0.1**, N = 2	**0.5**, N = 1	**0.6**, N = 1	**0.7**, N = 1	**0.8**, N =
bpm_att	2.39 (2.59)	6.00 (1.41)	7.00 (NA)	3.00 (NA)	7.00 (NA)	3.00 (1.73)
Missing	11	0	0	0	0	0
bpm_att_p	1.70 (2.10)	1.00 (1.41)	5.00 (NA)	NA (NA)	8.00 (NA)	4.00 (2.00)
Missing	11	0	0	1	0	0
bpm_ext	2.35 (2.10)	4.00 (0.00)	4.00 (NA)	4.00 (NA)	6.00 (NA)	2.33 (1.53)
Missing	11	0	0	0	0	0
bpm_ext_p	1.04 (1.74)	1.00 (0.00)	3.00 (NA)	11.00 (NA)	5.00 (NA)	4.33 (2.52)
Missing	11	0	0	0	0	0
bpm_int	2.09 (2.24)	3.50 (2.12)	3.00 (NA)	NA (NA)	8.00 (NA)	1.00 (1.00)
Missing	12	0	0	1	0	0
bpm_int_p	1.48 (1.85)	4.00 (5.66)	4.00 (NA)	7.00 (NA)	6.00 (NA)	3.67 (4.62)
Missing	9	0	0	0	0	0

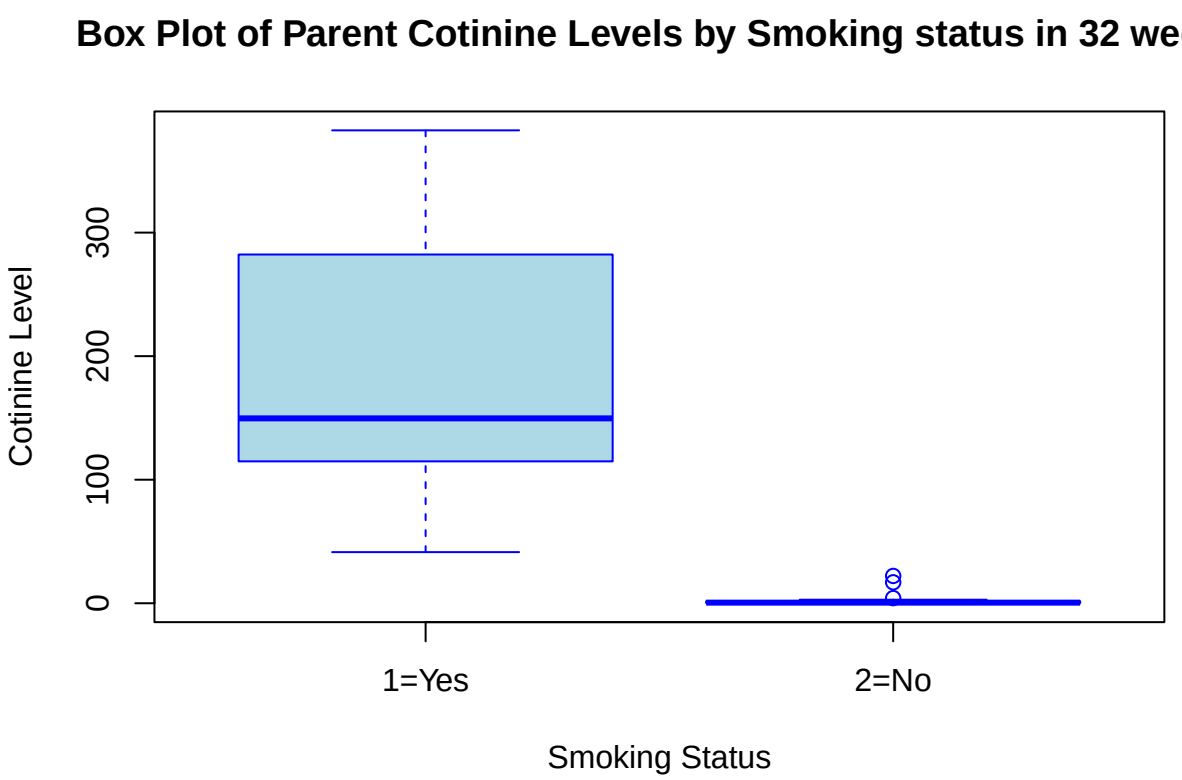
First we look at the statistics from the swan inattentive and swan hyperactive variable. It shows the median score and its confidence interval but this time we'll be evaluating these effects from the average ETS score from follow-up .For swan inattentive at avergae score of 0.1 we see the median is 13 with a lower bound of 10 and upper bound of 15. We see that for the inattentive variable that average ETS for 0 has a range between 7 and 13 with a median of 9 and average score 0.9 has an interval of (4,4) with a median of 4 since there is only one person to analyze.The p-value is 0.067, suggests that there is no statistically significant difference in the "swan_inattentive" variable between the average scores. For swan hyperactive the p-value as well did not prove statistical significance with a p-value of 0.2 and the different medians and confidence intervals can be seen in the summary table above.

I then observed the Emotion Regulation Questionairre responses between the average ETS follow-up scores and erq_exp and exp_cog have non-sginificant p-value both greater than 0.13.Where the average ETS score is 1 for erq_exp we see the median is 3.75 with a narrow CI with a lower bound of the median value 3.50 and upper bound of 4.00. We see that for the erq_cog variable that average ETS score for 0 has a wider range between 2.58,4.00 with a median of 3.17.

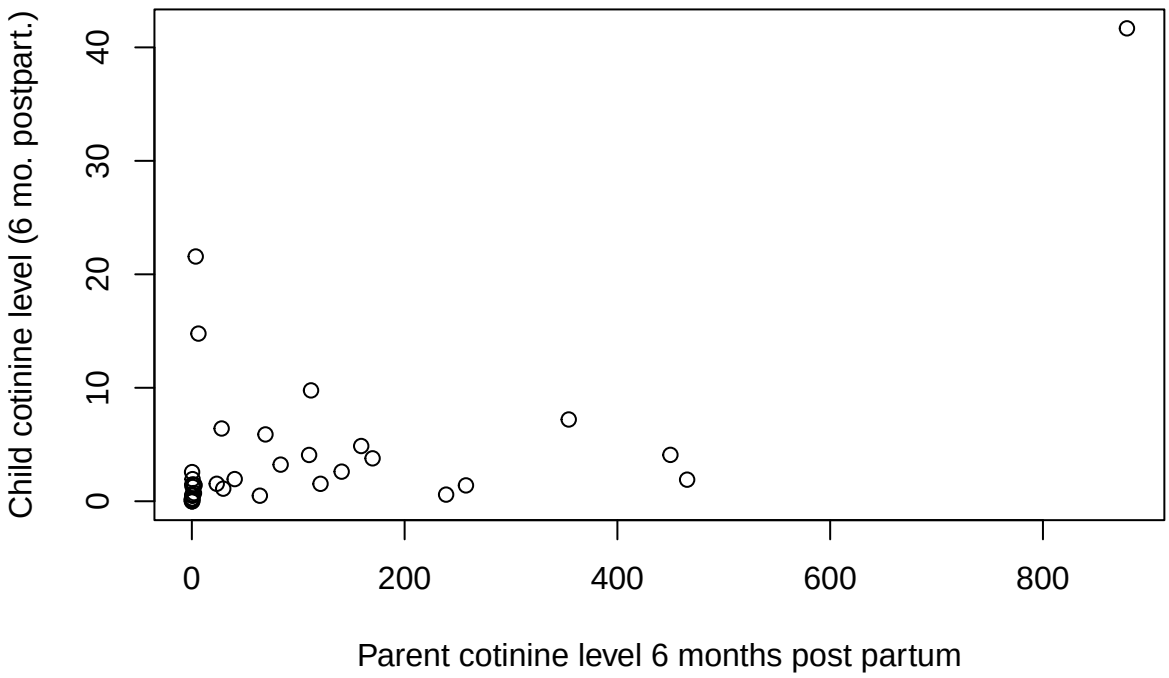
For the Brief Problem Monitor responses only none of the variables showed statistical significance but bpm_ext_p did also show statistical signifcance with a p-value of 0.054 .

In conclusion with the effects of SDP and ETS on the externalizing,substance use,and self-regulation variables there were not too many significant relationships we got results from. And a lot of these variables could be significant or on the cusp to reflect how smoking during pregnancy or exposure to tobacco smoke can have an effect on increasing substance use and poorer self-regulation if there was more data available and less missingess answering HYP1A and HYP1b.

```
#Plots of Cotimean measurements
cotimean_box<-boxplot(cotimean_34wk ~ mom_smoke_32wk,
        data = new_df,
        main = "Box Plot of Parent Cotinine Levels by Smoking status in 32 weeks",
        xlab = "Smoking Status",
        ylab = "Cotinine Level",
        col = c("lightblue", "lightgreen", "lightpink"),
        border = "blue")
```




```
contimean_post<-plot(new_df$cotimean_pp6mo, new_df$cotimean_pp6mo_baby,
xlab = "Parent cotinine level 6 months post partum",
ylab = "Child cotinine level (6 mo. postpart.)")
```



I wanted to look at the distribution of the contimean score between the mothers smoking status in her last trimester. We see that when answered Yes for their smoking status there is more of range and a higher average which would be predicted since continine does measure and test screens for nictoine which is present in the yes. This shows that the mothers answered truthfully about their smoking status in the last trimester. The same can be said for no with its small to none range and the couple of outliers in which can be accounted for fi they smoked recently prior to the testing in this last trimester of pregnancy.

The child continine levels at 6 months post-partum compared the mother's shows a high concentration at when the mother's measurement is 0 the child's is also 0. Even when the parents have high measurements of continine it is good to see that the tobacco screening is not reflected in child's urine meaning that it has not been passed onto the child.

AIM 2: Explore links between self-regulation at baseline and substance and externalizing at 6- and 12-month follow-ups.

Aim 2 involves investigating the connections between self-regulation measured at the baseline and substance use as well as externalizing behavior at the 6- and 12-month follow-ups. Given that our dataset contains data exclusively from the baseline assessment, I want to provide a overview of the baseline characteristics related to substance use, self-regulation, and externalizing behavior. Furthermore, I want to present a summary of these characteristics stratified by the cumulative (SDP) variable.

AIM3: Identify self-regulation problems that mediate the link between SDP/ETS exposure and level of, and change in, SU and EXT severity over time.