Exploratory Data Analysis Project: Effect of Tobacco Exposure on Child

Angel Zheng

Oct 2023 for PHP2550

Data Background

Tobacco exposure plays a significant role in behavioral and child developmental research, particularly concerning smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) post-delivery, both potentially linked to children's health and behavior throughout their lives. This report aims to explore the correlation between SDP and ETS with externalizing behaviors, self-regulation, and substance use.

The dataset utilized for this report stems from a follow-up study conducted by Dr. Micalizzi. This study focuses on a subset of adolescents (N=100) and their mothers, randomly selected for participation from a previous investigation centered on smoke avoidance intervention among low-income women to reduce SDP and ETS.

[1]The original study recruited approximately 800 pregnant individuals, all exposed to smoke (including current smokers, those who quit smoking on their own, or those exposed to secondhand smoke), expecting a single baby, and having access to a telephone and video player, were randomly assigned to either an experimental or control group. Those in the experimental group received personalized newsletters focused on quitting smoking (5 during pregnancy, 3 post-pregnancy), and custom-tailored videos based on behavioral surveys (3 during pregnancy, 2 post-pregnancy). Meanwhile, the control group received general newsletters and videos covering various aspects of healthy pregnancy.

Key areas of focus within this report include demographic information of both parents and children, variables examining their relationships, factors related to externalizing behaviors (EXT), aspects of self-regulation (SR), and variables pertaining to substance use (SU).

Externalizing behaviors encompass three main aspects: assessment via the Brief Problem Monitor (BPM) questionnaire, the SWAN rating, and evaluation for Autism diagnosis.

Regarding self-regulation, the study emphasizes emotion regulation, specifically assessed using the Emotion Regulation Questionnaire, focusing on Cognitive Reappraisal and Expressive Suppression.

Lastly, the dataset records information on children's usage of cigarettes, e-cigarettes, marijuana, and alcohol, representing primary and prevalent substance-use issues among adolescents.

Data Preprocessing

The processing began by filtering child data to select specific columns related to the child baseline arm. Demographic variables were chosen, excluding unnecessary columns and renaming certain variables for clarity. Subsequently, sections of the dataset pertaining to cigarette, e-cigarette, marijuana, and alcohol usage were modified to create summarized variables for further analysis.

The data manipulation continued by addressing the Brief Problem Monitor scoring and emotional regulation aspects, aggregating and selecting specific columns accordingly. Some sections of the dataset, such as physical

development scale, life stress assessments, and variables related to dysregulation, were dropped to simplify the analysis.

Similarly, alterations were made to the parent data, focusing on demographic variables and scoring related to parental monitoring, while excluding redundant or unrelated columns. Scoring for parental knowledge, child disclosure, parental solicitation, and control were derived based on specific columns. Sections concerning chaos, adult temperament, and stress were removed for the purpose of the research.

Upon thorough examination of the dataset, various anomalies were detected, ranging from formatting inconsistencies to substantial outliers that indicated potential errors. To address these discrepancies, several corrective measures were undertaken during the data preprocessing phase. Instances of numeric entries, for instance '250,000,' were standardized to '250000' for consistency in representation. Outliers, such as responses exceeding the maximum allowable value, for example '40' where the maximum value should be '30,' were identified and marked as 'NA' to flag potential inaccuracies. In cases where entries represented ranges, such as '20-25,' the mean was computed and rounded to the nearest integer, ensuring coherence in the dataset. Additionally, meticulous checks were conducted on binary variables, unifying their format to ensure consistency across all records. These meticulous steps were pivotal in rectifying anomalies and establishing uniformity in the dataset for subsequent analyses.

Details of how those entries are modified are in code appendix.

Demographic Information, univariate

The demographic information is displayed below. The participants contain parents with race of white the most (61%), with no Asians in both parents and children. There is a relative balance in Hispanic/Latino ratios. However, with the imbalance in race, it is worth noting that the study might have limited generalizability.

Table 1: Ethnicity and Race Demographic

Variable	**Child**, N = 49	**Parent**, N = 49
ethnicity		
Hispanic or Latino	15 (41%)	13 (32%)
None Hispanic nor Latino	21 (57%)	28 (68%)
Prefer not to answer	1 (2.7%)	0 (0%)
race		
American Indian/Alaska Native	5 (14%)	4 (9.8%)
Black	12 (33%)	0 (0%)
Native Hawaiian/Pacific Islander	0 (0%)	6 (15%)
Others	5 (14%)	6 (15%)
White	14 (39%)	25 (61%)

[1] n (%)

Table 2 shows the employment status, highest education, and child sex. Most parents have full-time job and above-highschool degree (`pedu`> 1). The mean and median of parent income are observably different, indicating that the income is skewed towards lower side with more parents with lower income. The histogram of parent and child age is also being looked at in Figure 1, but it does not seem problematic.

Table 2: Socioeconomic and Child Sex Demographic

Variable	**N = 49**
employ	
0	12 (29%)
1	7 (17%)
2	22 (54%)
pedu	

Table 2: Socioeconomic and Child Sex Demographic *(continued)*

Variable	**N = 49**
0	3 (7.3%)
1	3 (7.3%)
2	5 (12%)
3	15 (37%)
4	3 (7.3%)
5	10 (24%)
6	2 (4.9%)
income	
Mean (SD)	63,138 (59,885)
Median (IQR)	46,848 (20,000, 70,000)
page	
Mean (SD)	38 (4)
Median (IQR)	37 (35, 39)
tage	
12	8 (22%)
13	10 (27%)
14	9 (24%)
15	8 (22%)
16	2 (5.4%)
tsex	13 (36%)
[1] n (%)	

Missing Pattern

This dataset is problematic in extent of missingness. Below is the table showing percentage missingness for the variables, and it can be seen that `num_30` of all substances is majority missing. This is due to that this variable is only collected if the child reported "yes" to the previous variable. Therefore, the `num_30` variables are excluded from this analysis; instead, we use the previous binary responses to substances use.

`mom_smoke_pp1` and `mom_smoke_pp2` also have a large proportion missing. Looking at the data it could be due to loss of followup after delivery of baby. Therefore, for uniformity, we focus on the latter smoke exposure report instead of the followups of self-report mom smoke for postnatal tobacco exposure.

Table 3: Missingness Summary

Variable Name	Variable	Number of Missing
num_cigs_30	48	97.959184
num_e_cigs_30	47	95.918367
num_mj_30	46	93.877551
num_alc_30	45	91.836735
mom_smoke_pp1	39	79.591837
childasd	28	57.142857
mom_smoke_pp2	20	40.816327
pmq_parental_control	16	32.653061
ppmq_parental_solicitation	15	30.612245
swan_hyperactive	14	28.571429
bpm_int	14	28.571429
pmq_parental_knowledge	14	28.571429
pmq_parental_solicitation	14	28.571429
bpm_att_p	13	26.530612
tsex	13	26.530612
alc_ever	13	26.530612
erq_cog	13	26.530612

Table 3: Missingness Summary *(continued)*

Variable Name	Variable	Number of Missing
erq_exp	13	26.530612
pmq_child_disclosure	13	26.530612
income	12	24.489796
bpm_ext_p	12	24.489796
ppmq_parental_knowledge	12	24.489796
ppmq_child_disclosure	12	24.489796
ppmq_parental_control	12	24.489796
tage	12	24.489796
language	12	24.489796
tethnic	12	24.489796
cig_ever	12	24.489796
e_cig_ever	12	24.489796
mj_ever	12	24.489796
bpm_att	12	24.489796
bpm_ext	12	24.489796
nidapres	11	22.448980
momcig	11	22.448980
mom_numcig	11	22.448980
cotimean_34wk	11	22.448980
cotimean_pp6mo_baby	11	22.448980
cotimean_pp6mo	11	22.448980
smoke_exposure_3yr	11	22.448980
smoke_exposure_4yr	11	22.448980
bpm_att_a	11	22.448980
bpm_ext_a	11	22.448980
nidaalc	10	20.408163
nidatob	10	20.408163
nidaill	10	20.408163
swan_inattentive	10	20.408163
bpm_int_p	10	20.408163
smoke_exposure_6mo	10	20.408163
smoke_exposure_12mo	10	20.408163
smoke_exposure_2yr	10	20.408163
smoke_exposure_5yr	10	20.408163
bpm_int_a	10	20.408163
erq_cog_a	10	20.408163
erq_exp_a	10	20.408163
mom_smoke_32wk	9	18.367347
mom_smoke_pp6mo	9	18.367347
page	8	16.326531
psex	8	16.326531
plang	8	16.326531
pethnic	8	16.326531
employ	8	16.326531
pedu	8	16.326531
mom_smoke_22wk	7	14.285714
mom_smoke_pp12wk	7	14.285714
mom_smoke_16wk	1	2.040816

Prenatal and Postnatal Smoking

In the below two tables, it can be seen that at each timepoint for both prenatal and postnatal period, the mothers who reported smoking or smoking exposure of children are the relative minority (around 30%).

Table 4: Prenatal Tobacco Exposure (Smoking During Pregnancy)

Variable	**N = 49**
mom_smoke_16wk	12 (25%)
mom_smoke_22wk	13 (31%)
mom_smoke_32wk	10 (25%)
cotinine	
0	26 (68%)
1	2 (5.3%)
2	10 (26%)
prenatal_score	
0	21 (60%)
1	3 (8.6%)
2	1 (2.9%)
5	10 (29%)

[1] n (%)

Table 5: Postnatal Tobacco Exposure (Environmental Tobacco Smoke)

Variable	**N = 49**
smoke_exposure_6mo	10 (26%)
smoke_exposure_12mo	9 (23%)
smoke_exposure_2yr	11 (28%)
smoke_exposure_3yr	11 (29%)
smoke_exposure_4yr	10 (26%)
smoke_exposure_5yr	10 (26%)
cotinine	
0	19 (50%)
1	3 (7.9%)
2	16 (42%)
postnatal_score	
0	15 (50%)
1	3 (10%)
2	4 (13%)
3	1 (3.3%)
5	2 (6.7%)
6	2 (6.7%)
7	1 (3.3%)
8	2 (6.7%)

[1] n (%)

This figure shows an relationship between SDP summary scores and ETS summmary scores, which are computed by adding the smoking behavior and cotinine levels. [2]Cotinine are categorized into 3 levels: nonsmoker (0) with cotinine < 10, passive-smoker (1) with cotinine < 30, and active smoker (2) with cotinine > 30 ng/mL. In the bar chart, it is seen that at lower ETS scores, there are more subjects with low SDP scores; and at higher ETS scores, there are mostly high SDP scores subjects. This demonstrates that SDP and ETS are related to each other positively, and is also confirmed by their correlation value $r = 0.683$ which is nearly highly correlated with each other.

Externalizing Behaviors vs. SDP/ETS

Recall that autism diagnosis is a part of externalizing behaviors, but it won't be included because as only 1 child has been diagnosed and 1 child is suspected, while all others no autism or missing response. In this case, including autism would not be helpful to the analysis.

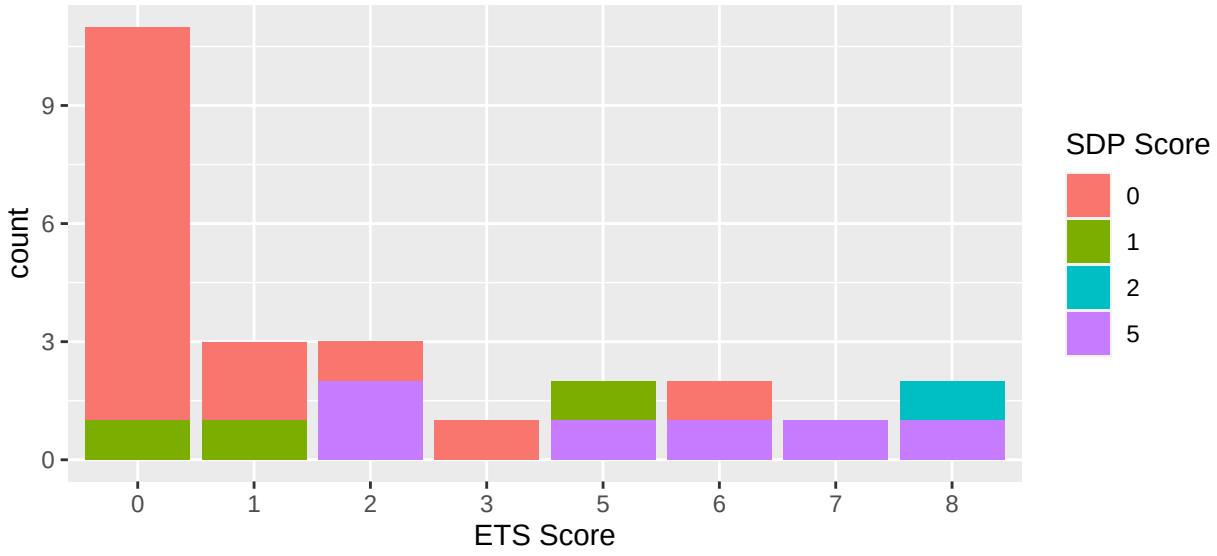

Figure 1: The Relationship Between SDP and ETS Summary Scores

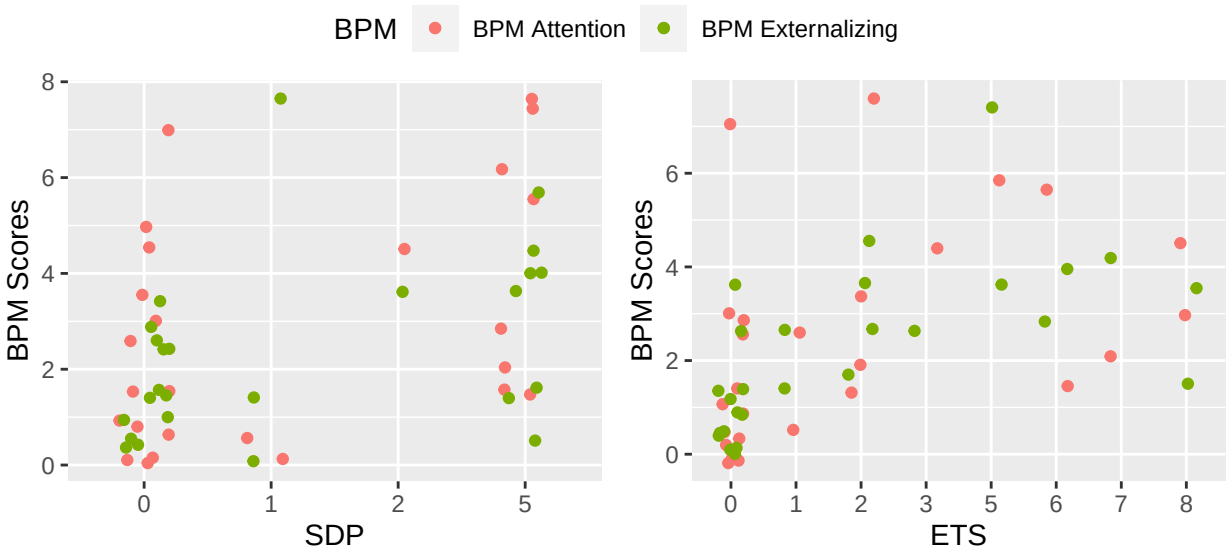

Figure 2: Effect of SDP/ETS on BPM Scores

BPM and SWAN questionnaires are plotted with different colors.

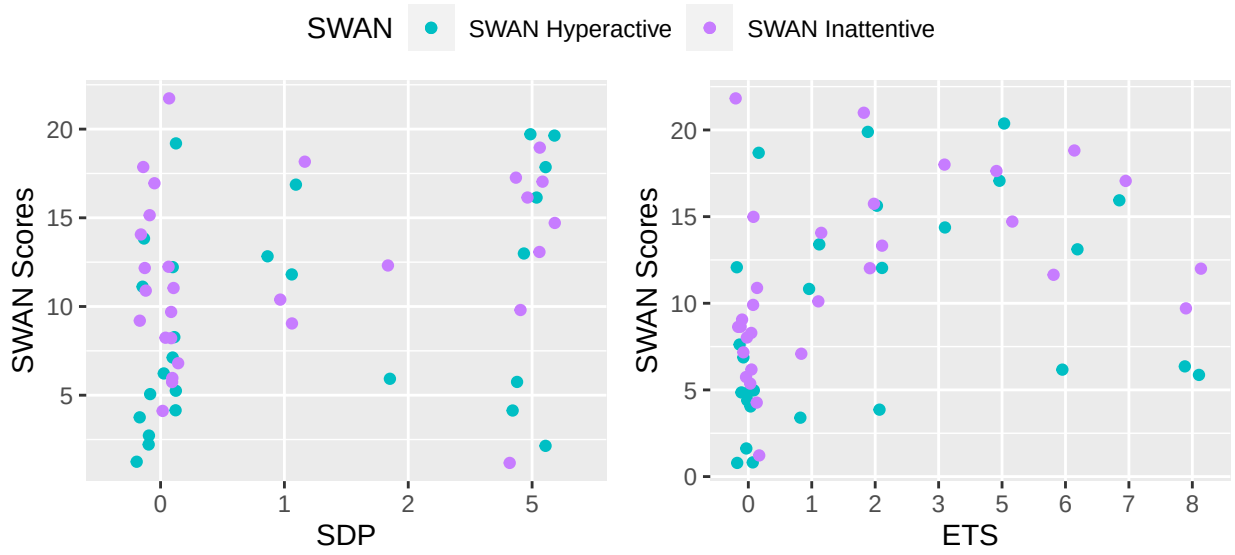

Figure 3: Effect of SDP/ETS on SWAN Scores

The figures display the relationship between SDP/ETS and the questionnaire response on externalizing behaviors, specifically the BPM and SWAN scores. Regardless of the limited data points, it can still be seen that as the severity of SDP and ETS increases, the questionnaire responses tend to cluster at a higher score. This suggests a sign that SDP and ETS is associated with increased problem in externalizing behaviors.

In this section, some correlations are also noted:

`cor(Attention BPM Score of Child, Attention BPM Score of Parent)` $r = 0.571$

`cor(Severity of Postnatal Smoke Exposure, Child-Rated Parental Knowledge)` $r = -0.189$

`cor(Severity of Postnatal Smoke Exposure, Child-Rated Child Disclosure)` $r = -0.290$

BPM score of child and parent are moderately correlated ($r = 0.571$), which adds a confounding to the associations between attention BPM scores and SDP/ETS.

Also, severtiy of postnatal ETS is negatively correlated to both child-rated parental knowledge and child disclosure, which draws a relationship between ETS and irresponsibility of parent (although the correlation is not strong). This provides a counter-argument to the validity of SWAN rating, which was by the parent to the child. In the families with more severe postnatal ETS, the parents might not be responsible enough to know if their children are showing symptoms for ADHD.

Self Emotion Regulation vs. SDP/ETS

In this section, because the outcome is still responses to questionnaires, we take similar approach as in the last section. Boxplots and scatter points here show ERQ scores on cognitive reappraisal and expressive suppression do not show a clear relationship with SDP or ETS. The median is varying, and the range is also drastically different. Boxplots are included to demonstrate that at some levels of SDP and ETS, there were extremely limited entries which could be the reason that this relationship is difficult to examine.

In this section, some correlations are also noted:

`cor(ERQ Score of Child Cognitive Reappraisal, ERQ Score of Child Expressive Suppression)` $r = 0.119$

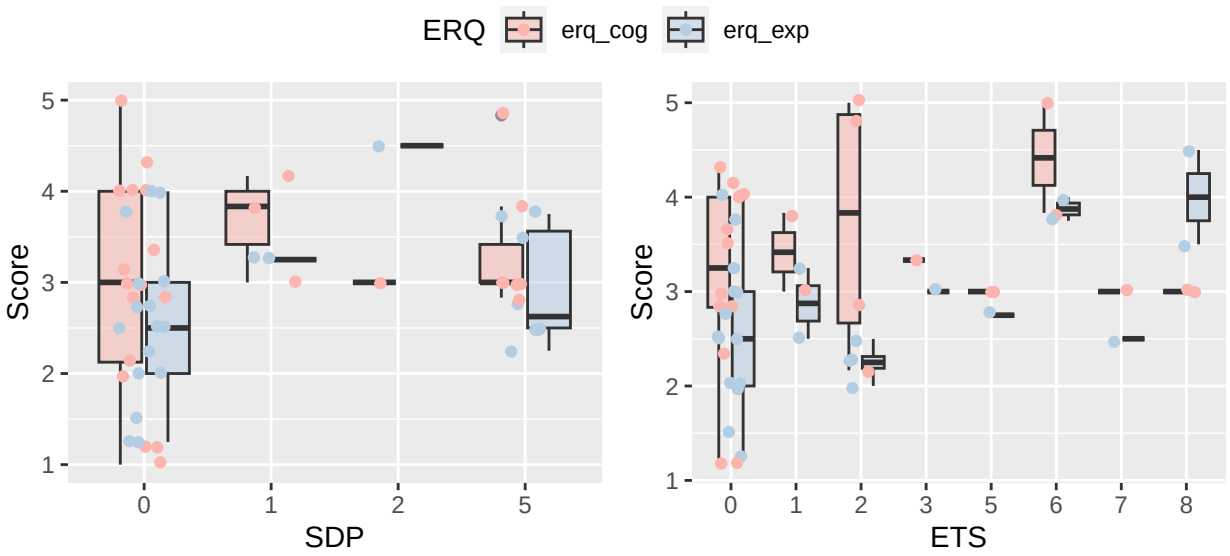

Figure 4: Effect of SDP/ETS on Self Emotion Regulation

`cor(ERQ Score of Child Cognitive Reappraisal, ERQ Score of Parent Cognitive Reappraisal)` $r = 0.090$

`cor(ERQ Score of Child Expressive Suppression, ERQ Score of Parent Expressive Suppression)` $r = 0.258$

The inter-correlation between the two aspects of emotion control is weak. The correlations between parent's emotion control and child's are also weak.

Substance-use vs. Prenatal and Postnatal Period Tobacco Exposure

There is relatively less available data to analyze in this substance-use section. Relating to the age of children being included, it is reasonable that some of them never tried any of the substance.

This section uses bar charts with proportion of children using certain substance. Color of the bar indicates the smoking or exposure status. In the prenatal barplots, for all timepoints, there is a higher proportion of substance-use child for all types of substance if the mother self-reported as smoker (Missing bar due to proportion = 0, no child with subtance use in that group). However, in the postnatal barplots, there is higher proportion of cigarettes and marijuanause for children with non-smoker mother.

Initial Regression

In this section, regressions for the outcome variables in each aspect (EXT, SR, SU) -> `pre/post timepoints*smoke status` are generated.

In the externalizing part, the smoke status in prenatal period is statistically significant $p = 0.040$ to the BPM score of attention problems Also, the smoke status in postnatal period is statistically significant $p = 0.011$ to the BPM score of externalizing problems. There is no shown significance for the emotion control outcomes. And possibly due to the limited availability of substance use outcomes, the smoke status and timepoints all have small p-values in pre&postnatal periods.

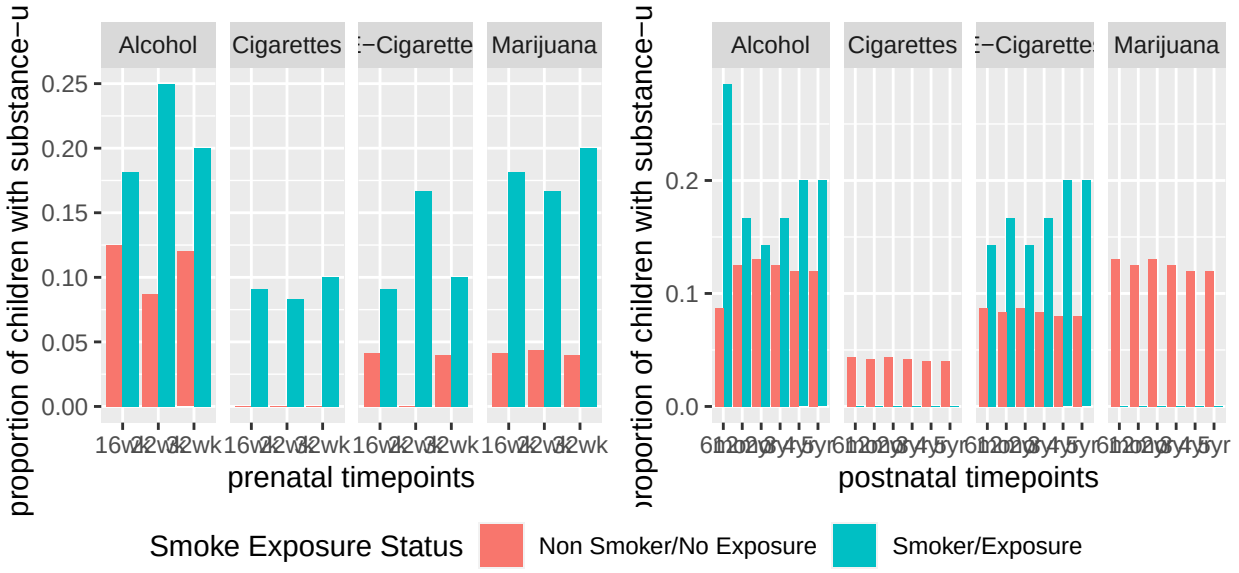

Figure 5: Substance Use vs. Prenatal Timepoints by Substance Type

```
# example chunck for initial regression
summary(lm(bpm_att_mean ~ timepoint*smoke, data=ext.pre)) # smoke status significant
summary(lm(bpm_ext_mean ~ timepoint*smoke, data=ext.post)) # smoke status significant
```

Conclusion

After conducting an in-depth Exploratory Data Analysis, the impact of SDP/ETS on adolescent behavior, externalizing tendencies, self-regulation, and substance use was thoroughly investigated. The evidence suggests a noticeable correlation between the presence of SDP/ETS during prenatal or postnatal periods and the onset of externalizing problems in children. These problems encompass attention issues and symptoms resembling ADHD. Moreover, a higher incidence of substance use problems was observed among children with mothers who reported SDP. However, clear evidence regarding the effects of SDP/ETS on emotion regulation remains elusive.

This dataset boasts several strengths and limitations. It encompasses a wide array of variables, including demographic information and multiple measurements of behavioral aspects. This comprehensive range of variables opens up diverse possibilities for data analysis and offers transparency in subject traits for future researchers seeking replication.

Nevertheless, a drawback of this extensive variable range is the presence of missing data. Loss of follow-up and missing baseline variables contribute to this missingness, reducing the dataset's statistical power. Imputation methods aren't suitable due to the small sample size. Furthermore, the limited number of instances with reported SDP/ETS within this small sample size restricts in-depth analysis, potentially leading to chance effects.

Certain variables, notably maternal responses regarding smoking during pregnancy and child-reported substance use, rely on self-reporting and may be influenced by social desirability bias. Subjects might choose not to disclose truthfully, aligning their responses with societal norms.

In conclusion, while the plots in this report serve as a preliminary exploration, a more detailed analysis is advisable before arriving at a conclusive scientific inference. The findings presented here should be viewed as indicative and require further scrutiny.

6 Reference

[1] Risica, P. M., Gavarkovs, A., Parker, D. R., Jennings, E., & Phipps, M. (2017). A tailored video intervention to reduce smoking and environmental tobacco exposure during and after pregnancy: Rationale, design and methods of Baby's Breath. Contemporary clinical trials, 52, 1–9. https://doi.org/10.1016/j.cct.2016.10.010

[2] University of Rochester Medical Center. (n.d.). Nicotine & Cotinine. Retrieved from https://www.urmc.rochester.edu/encyclopedia/content.aspx?contentid=nicotine_cotinine&contenttypeid=167#:~:text=Cotinine%20levels%20in%20a%20nonsmoker,more%20than%20500%20ng%2FmL.

7 Code Appendix: All code for this report

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning=FALSE)

library(tidyverse) # for data manimulation
library(ggmice) # for missing pattern
library(naniar) # for missing summary
library(kableExtra) # for neat output
library(gtsummary) # for neat summary table
library(ggpubr) # for combining ggplots

tobacco_exposure <-  read.csv("C:/ANGEL/Brown 22-24/23Fall/PHP2550_pda/datasets/project1.csv")

#colnames(tobacco_exposure)

# make "250,000" as "250000" to uniform numeric format
tobacco_exposure$income[which(tobacco_exposure$income == "250, 000")] <- 250000
tobacco_exposure$income <- as.integer(tobacco_exposure$income)

# Null entry with 40 days of smoking in 30 days
tobacco_exposure$momcig[which(tobacco_exposure$momcig == 40)] <- NA

# Null entry with unreasonable response
tobacco_exposure$mom_numcig[which(tobacco_exposure$mom_numcig == 44989)] <- NA

# convert text response to numeric
tobacco_exposure$mom_numcig[which(tobacco_exposure$mom_numcig == "2 black and miles a day")] <- 2
tobacco_exposure$mom_numcig[which(tobacco_exposure$mom_numcig == "20-25")] <- 23
tobacco_exposure$mom_numcig[which(tobacco_exposure$mom_numcig == "None")] <- 0
tobacco_exposure$mom_numcig <- as.integer(tobacco_exposure$mom_numcig)

# extract numeric part from string responses
string_to_binary <- function(string_var) {
  string_num <- ifelse(string_var == "", NA, # NA if blank
                       as.integer(gsub("([0-9]+).*$", "\\1", string_var)))
  string_binary <- replace(string_num, string_num == 2, 0)
  return(string_binary)
}

tobacco_exposure[, c("mom_smoke_16wk", "mom_smoke_22wk", "mom_smoke_32wk",
                     "mom_smoke_pp1", "mom_smoke_pp2",
                     "mom_smoke_pp12wk", "mom_smoke_pp6mo")] <- lapply(
                       tobacco_exposure[, c("mom_smoke_16wk", "mom_smoke_22wk", "mom_smoke_32wk",
                                            "mom_smoke_pp1", "mom_smoke_pp2",
                                            "mom_smoke_pp12wk", "mom_smoke_pp6mo")],
                       string_to_binary)

# correct 0 in SWAN response to NA
tobacco_exposure$swan_inattentive[which(tobacco_exposure$swan_inattentive == 0)] <- NA
tobacco_exposure$swan_hyperactive[which(tobacco_exposure$swan_hyperactive == 0)] <- NA

# create a datafrmae for reporting race and ethnicity
race_ethnicity <-
```

```
  tobacco_exposure %>%
  mutate(
    parent_ethnicity = case_when(
      pethnic == 1 ~ "Hispanic or Latino",
      pethnic == 0 ~ "None Hispanic nor Latino",
      pethnic == 2 ~ "Prefer not to answer"),
    parent_race = case_when(
      paian == 1 ~ "American Indian/Alaska Native",
      pasian == 1 ~ "Asian",
      pnhpi == 1 ~ "Native Hawaiian/Pacific Islander",
      pblack == 1 ~ "Black",
      pwhite == 1 ~ "White",
      prace_other == 1 ~ "Others"),
    child_ethnicity = case_when(
      tethnic == 1 ~ "Hispanic or Latino",
      tethnic == 0 ~ "None Hispanic nor Latino",
      tethnic == 2 ~ "Prefer not to answer"),
    child_race = case_when(
      taian == 1 ~ "American Indian/Alaska Native",
      tasian == 1 ~ "Asian",
      tnhpi == 1 ~ "Native Hawaiian/Pacific Islander",
      tblack == 1 ~ "Black",
      twhite == 1 ~ "White",
      trace_other == 1 ~ "Others")
  ) %>%
  select(parent_ethnicity, parent_race, child_ethnicity, child_race)

# Table 1: summary of race and ethnicity
data.frame(parent_child = rep(c("Parent", "Child"), each = 49),
           ethnicity = c(race_ethnicity$parent_ethnicity, race_ethnicity$child_ethnicity),
           race = c(race_ethnicity$parent_race, race_ethnicity$child_race)) %>%
  tbl_summary(by = parent_child, missing = "no") %>%
  modify_header(label ~ "**Variable**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Ethnicity and Race Demographic",
                 longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

# Table 2: Summary of socioeconomic and child sex
tobacco_exposure %>%
  select(c(12:14,2,52:53)) %>%
  tbl_summary(missing="no",
              type = all_continuous() ~ "continuous2",
              statistic = list(all_continuous() ~ c("{mean} ({sd})", "{median} ({p25}, {p75})")))  %>%
  modify_header(label ~ "**Variable**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Socioeconomic and Child Sex Demographic",
                 longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

# Missingness summary for all variables
```

```r
missingness_te <-
  tobacco_exposure %>%
  miss_var_summary() %>%
  filter(n_miss > 0)

colnames(missingness_te) <- c("Variable Name", "Variable", "Number of Missing", "Percentage % of Missing

missingness_te %>%
  kable(booktabs = TRUE, caption = "Missingness Summary", longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position"))

# convert cotinine level to categorical
cotinine_pre <-
  ifelse(tobacco_exposure$cotimean_34wk < 10, 0,
       ifelse(tobacco_exposure$cotimean_34wk < 30, 1, 2))

cotinine_post <-
  ifelse(tobacco_exposure$cotimean_pp6mo < 10, 0,
       ifelse(tobacco_exposure$cotimean_pp6mo < 30, 1, 2))

# a dataframe for prenatal responses to self-report smoking behavior
prenatal_severity <-
  tobacco_exposure[,c("parent_id", "mom_smoke_16wk", "mom_smoke_22wk", "mom_smoke_32wk")] %>%
  mutate(cotinine = cotinine_pre,
         prenatal_score = as.factor(cotinine + mom_smoke_16wk + mom_smoke_22wk + mom_smoke_32wk))

# a dataframe for postnatal responses to tobacco exposure
postnatal_severity <-
tobacco_exposure[,c("parent_id", "smoke_exposure_6mo", "smoke_exposure_12mo",
                    "smoke_exposure_2yr", "smoke_exposure_3yr", "smoke_exposure_4yr", "smoke_exposure_5y
  mutate(cotinine = cotinine_post,
         postnatal_score = as.factor(cotinine + smoke_exposure_6mo + smoke_exposure_12mo +
                    smoke_exposure_2yr + smoke_exposure_3yr + smoke_exposure_4yr + smoke_exposure_5yr))
# Table summary of prenatal smoking
tbl_summary(prenatal_severity[,-1], missing = "no") %>%
  modify_header(label ~ "**Variable**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Prenatal Tobacco Exposure (Smoking During Pregnancy)",
                 longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

# Table 5: summary of postnatal smoking
tbl_summary(postnatal_severity[,-1], missing = "no") %>%
  modify_header(label ~ "**Variable**") %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Postnatal Tobacco Exposure (Environmental Tobacco Smoke)",
                 longtable = TRUE) %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

SDP_ETS <- na.omit(merge(prenatal_severity, postnatal_severity, by = "parent_id"))
```

```r
ggplot(SDP_ETS, aes(postnatal_score)) +
  geom_bar(aes(fill=prenatal_score)) +
  xlab("ETS Score") +
  guides(fill=guide_legend(title="SDP Score"))

#cor(as.numeric(SDP_ETS$prenatal_score), as.numeric(SDP_ETS$postnatal_score))

# a subset of dataframe selecting variables of externalizing
externalizing <-
  tobacco_exposure %>%
  select("parent_id", "bpm_att", "bpm_att_p", "bpm_ext", "bpm_ext_p", "swan_hyperactive", "swan_inatten
  mutate(bpm_att_mean = (bpm_att + bpm_att_p)/2) %>% # child attention score from self and parent
  mutate(bpm_ext_mean = (bpm_ext + bpm_ext_p)/2) %>% # child externalizing score from self and parent
  merge(prenatal_severity, "parent_id") %>% # combine with prenatal exposure
  merge(postnatal_severity, "parent_id")  # combine with postnatal exposure

# autism data distribution
#table(externalizing$childasd, useNA = "always") # not helpful to include autism data

bpm_colors <- c("BPM Attention" = "#F8766D", "BPM Externalizing" = "#7CAE00")

BPM_SDP <-
  ggplot(subset(externalizing, !is.na(prenatal_score)), aes(prenatal_score)) +
  geom_jitter(aes(y=bpm_att_mean, color = "BPM Attention"), width = 0.2) +
  geom_jitter(aes(y=bpm_ext_mean, color = "BPM Externalizing"), width = 0.2) +
  labs(x = "SDP", y = "BPM Scores", color = "BPM") +
  scale_color_manual(values = bpm_colors)

BPM_ETS <-
  ggplot(subset(externalizing, !is.na(postnatal_score)), aes(postnatal_score)) +
  geom_jitter(aes(y=bpm_att_mean, color = "BPM Attention"), width = 0.2) +
  geom_jitter(aes(y=bpm_ext_mean, color = "BPM Externalizing"), width = 0.2) +
  labs(x = "ETS", y = "BPM Scores", color = "BPM") +
  scale_color_manual(values = bpm_colors)

ggarrange(BPM_SDP, BPM_ETS, common.legend = TRUE)

swan_colors <- c("SWAN Hyperactive" = "#00BFC4", "SWAN Inattentive" = "#C77CFF")

SWAN_SDP <-
  ggplot(subset(externalizing, !is.na(prenatal_score)), aes(prenatal_score)) +
  geom_jitter(aes(y=swan_hyperactive, color = "SWAN Hyperactive"), width = 0.2) +
  geom_jitter(aes(y=swan_inattentive, color = "SWAN Inattentive"), width = 0.2) +
  labs(x = "SDP", y = "SWAN Scores", color = "SWAN") +
  scale_color_manual(values = swan_colors)

SWAN_ETS <-
  ggplot(subset(externalizing, !is.na(postnatal_score)), aes(postnatal_score)) +
  geom_jitter(aes(y=swan_hyperactive, color = "SWAN Hyperactive"), width = 0.2) +
  geom_jitter(aes(y=swan_inattentive, color = "SWAN Inattentive"), width = 0.2) +
  labs(x = "ETS", y = "SWAN Scores", color = "SWAN") +
  scale_color_manual(values = swan_colors)
```

```r
ggarrange(SWAN_SDP, SWAN_ETS, common.legend = TRUE)

# some correlations being tested
cor(externalizing$bpm_att_mean, tobacco_exposure$bpm_att_a, use = "pairwise.complete.obs")
cor(externalizing$bpm_ext_mean, tobacco_exposure$bpm_ext_a, use = "pairwise.complete.obs")
cor(postnatal_severity$postnatal_score, tobacco_exposure$pmq_parental_knowledge, use = "pairwise.comple
cor(postnatal_severity$postnatal_score, tobacco_exposure$pmq_child_disclosure, use = "pairwise.complete

# prepare format for plotting
erq.pre <- prenatal_severity %>%
  merge(tobacco_exposure[,c("parent_id", "erq_cog", "erq_exp")], "parent_id") %>%
  select(-prenatal_score) %>%
  pivot_longer(cols = starts_with("mom_smoke_"), names_to = "timepoint", values_to = "smoke") %>%
  subset(!is.na(smoke))

erq.pre$timepoint <- sub("^mom_smoke_(.*)$", "\\1", erq.pre$timepoint)

erq.pre <-
  prenatal_severity %>%
  merge(tobacco_exposure[,c("parent_id", "erq_cog", "erq_exp")], "parent_id") %>%
  select(6:8) %>%
  pivot_longer(cols=!prenatal_score, names_to = "ERQ", values_to = "Score")

erq.post <-
  postnatal_severity %>%
  merge(tobacco_exposure[,c("parent_id", "erq_cog", "erq_exp")], "parent_id") %>%
  select(9:11) %>%
  pivot_longer(cols=!postnatal_score, names_to = "ERQ", values_to = "Score")

ERQ_SDP <-
  ggplot(aes(x=prenatal_score, y=Score), data=subset(erq.pre, !is.na(prenatal_score))) +
  geom_boxplot(aes(fill=ERQ), alpha=0.5) +
  geom_jitter(aes(color=ERQ), width=0.2) +
  scale_fill_brewer(palette = "Pastel1") +
  scale_color_brewer(palette = "Pastel1") +
  labs(x="SDP")

ERQ_ETS <-
  ggplot(aes(x=postnatal_score, y=Score), data=subset(erq.post, !is.na(postnatal_score))) +
  geom_boxplot(aes(fill=ERQ), alpha=0.5) +
  geom_jitter(aes(color=ERQ), width=0.2) +
  scale_fill_brewer(palette = "Pastel1") +
  scale_color_brewer(palette = "Pastel1") +
  labs(x="ETS")

ggarrange(ERQ_SDP, ERQ_ETS, common.legend = TRUE)

cor(tobacco_exposure$erq_cog, tobacco_exposure$erq_exp, use = "pairwise.complete.obs")
cor(tobacco_exposure$erq_cog, tobacco_exposure$erq_cog_a, use = "pairwise.complete.obs")
cor(tobacco_exposure$erq_exp, tobacco_exposure$erq_exp_a, use = "pairwise.complete.obs")

# prepare format for plotting
substance.pre <-
```

```
  tobacco_exposure %>%
  select("parent_id", "cig_ever", "e_cig_ever", "mj_ever", "alc_ever") %>%
  merge(prenatal_severity, "parent_id") %>%
  pivot_longer(cols = starts_with("mom_smoke_"), names_to = "timepoint", values_to = "smoke") %>%
  subset(!is.na(prenatal_score)) %>%
  group_by(timepoint, smoke) %>%
  mutate(cig_prop = sum(cig_ever, na.rm = TRUE)/n(),
         e_cig_prop = sum(e_cig_ever, na.rm = TRUE)/n(),
         mj_prop = sum(mj_ever, na.rm = TRUE)/n(),
         alc_prop = sum(alc_ever, na.rm = TRUE)/n())

# modify timepoint format
substance.pre$timepoint <- sub("^mom_smoke_(.*)$", "\\1", substance.pre$timepoint)

# labels for facets
substance.labs <- c("Alcohol", "Cigarettes", "E-Cigarettes", "Marijuana")
names(substance.labs) <- c("alc_prop", "cig_prop", "e_cig_prop", "mj_prop")

# prepare format for plotting
substance.post <-
  tobacco_exposure %>%
  select("parent_id", "cig_ever", "e_cig_ever", "mj_ever", "alc_ever") %>%
  merge(postnatal_severity, "parent_id") %>%
  pivot_longer(cols = starts_with("smoke_exposure"), names_to = "timepoint", values_to = "smoke") %>%
  subset(!is.na(postnatal_score)) %>%
  group_by(timepoint, smoke) %>%
  mutate(cig_prop = sum(cig_ever, na.rm = TRUE)/n(),
         e_cig_prop = sum(e_cig_ever, na.rm = TRUE)/n(),
         mj_prop = sum(mj_ever, na.rm = TRUE)/n(),
         alc_prop = sum(alc_ever, na.rm = TRUE)/n())

# modify timepoint format
postnatal_order <- c("6mo", "12mo", "2yr", "3yr", "4yr", "5yr")
substance.post$timepoint <- factor(sub("^smoke_exposure_(.*)$", "\\1", substance.post$timepoint), levels

# bar plots for substance use vs. prenatal smoke status
substance_bar.pre <-
  substance.pre %>%
  select(timepoint, smoke, cig_prop:alc_prop) %>%
  unique() %>%
  gather(substance, prop, cig_prop:alc_prop) %>%
  ggplot(aes(x=timepoint, y=prop, fill=as.factor(smoke))) +
    geom_col(position="dodge") +
    facet_grid(. ~ substance, labeller = labeller(substance=substance.labs)) +
    xlab("prenatal timepoints") + ylab("proportion of children with substance-use") +
    scale_fill_discrete(name = "Smoke Exposure Status", labels = c("Non Smoker/No Exposure", "Smoker/Exp

# bar plots for substance use vs. postnatal smoke status
substance_bar.post <-
substance.post %>%
  select(timepoint, smoke, cig_prop:alc_prop) %>%
  unique() %>%
  gather(substance, prop, cig_prop:alc_prop) %>%
```

```
  ggplot(aes(x=timepoint, y=prop, fill=as.factor(smoke))) +
    geom_col(position="dodge") +
    facet_grid(. ~ substance, labeller = labeller(substance=substance.labs)) +
    xlab("postnatal timepoints") + ylab("proportion of children with substance-use")

# FSubstance Use vs. Prenatal Timepoints by Substance Type
ggarrange(substance_bar.pre, substance_bar.post, common.legend = TRUE, legend="bottom")

# substance use proportion vs. prenatal cotinine levels of parent
# not included in report, discarded
substance.pre %>%
  merge(tobacco_exposure[,c("parent_id", "cotimean_34wk")], "parent_id") %>%
  gather(substance, prop, cig_prop:alc_prop) %>%
  select(9:11) %>%
  unique() %>%
  ggplot(aes(x=cotimean_34wk, y=prop, color=substance)) +
    geom_point() + geom_smooth(method = lm, se = FALSE)

# substance use proportion vs. postnatal cotinine levels of baby
# not included in report, discarded
substance.post %>%
  merge(tobacco_exposure[,c("parent_id", "cotimean_pp6mo_baby")], "parent_id") %>%
  gather(substance, prop, cig_prop:alc_prop) %>%
  select(9:11) %>%
  unique() %>%
  ggplot(aes(x=cotimean_pp6mo_baby, y=prop, color=substance)) +
    geom_point() + geom_smooth(method = lm, se = FALSE)

# example chunck for initial regression
summary(lm(bpm_att_mean ~ timepoint*smoke, data=ext.pre)) # smoke status significant
summary(lm(bpm_ext_mean ~ timepoint*smoke, data=ext.post)) # smoke status significant

# Regressions to check p-values
summary(lm(bpm_att_mean ~ timepoint*smoke, data=ext.pre)) # smoke status significant
summary(lm(bpm_ext_mean ~ timepoint*smoke, data=ext.pre))
summary(lm(swan_hyperactive ~ timepoint*smoke, data=ext.pre))
summary(lm(swan_inattentive ~ timepoint*smoke, data=ext.pre))

summary(lm(bpm_att_mean ~ timepoint*smoke, data=ext.post))
summary(lm(bpm_ext_mean ~ timepoint*smoke, data=ext.post)) # smoke status significant
summary(lm(swan_hyperactive ~ timepoint*smoke, data=ext.post))
summary(lm(swan_inattentive ~ timepoint*smoke, data=ext.post))

summary(lm(erq_cog ~ timepoint*smoke, data=erq.pre))
summary(lm(erq_exp~ timepoint*smoke, data=erq.pre))

summary(lm(erq_cog ~ timepoint*smoke, data=erq.post))
summary(lm(erq_exp~ timepoint*smoke, data=erq.post))

summary(lm(cig_prop ~ timepoint*smoke, data=substance.pre)) # smoke and time trend significant
summary(lm(e_cig_prop ~ timepoint*smoke, data=substance.pre))
summary(lm(mj_prop ~ timepoint*smoke, data=substance.pre))
summary(lm(alc_prop ~ timepoint*smoke, data=substance.pre))
```

```r
summary(lm(cig_prop ~ timepoint*smoke, data=substance.post)) # smoke and time trend significant
summary(lm(e_cig_prop ~ timepoint*smoke, data=substance.post))
summary(lm(mj_prop ~ timepoint*smoke, data=substance.post))
summary(lm(alc_prop ~ timepoint*smoke, data=substance.post))
```