Investigating the Effects of Smoking During Pregnancy and Environmental Tobacco Smoking on Adolescent Self-Regulation and Externalizing Behavior

Wanyi Chen

Introduction

Background

Maternal SDP has been a major public health concern, and it is creating adverse consequences for offspring and costs to society. 7 to 15% of the U.S. infants born per year are exposed to SDP and the related health-care cost causes a $5 billion annual burden on the U.S. economy (Micalizzi). According to the statistics in Centers for Disease Control and Prevention (CDC), SDP can cause tissue damage in unborn baby and mothers are more likely to deliver their babies early. Preterm delivery is highly associated with lower birth weight which is a leading cause of death and disability among newborn babies. Furthermore, SDP-exposed children have increased rates of externalizaing behaviors such as substance use and attention-deficit, and self-regulatory problems such as executive function and emotion regulation. Therefore, the main objective of this project is to examine the effects of Smoke During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) on adolescent self-regulation, substance use, and externalizing through an Exploratory Data Analysis (EDA).

Data Source and Pre-processing

The data source is from a larger data set that is originally collected for a study on smoke avoidance intervention to reduce low-income women's smoking, and ETS exposure during pregnancy and children's exposure to ETS in the immediate postpartum period. For the purpose of this project, some preprocessing is performed to narrow down the scope of variables to focus on. In the child data set, information of demographic, child substance use, self-regulation problems, and externalizaing behaviors are extracted. In the parent data set, more information relating to SDP, ETS, income, employment, parental knowledge, child externalizing behaviors, mom and baby urine cotinine at different time points, SWAN rating, and child self-regulation problems are included. The complete data set has 98 adolescents and mothers that are randomly selected from the orginal larger data set.

Descriptive Statistics

First, we can briefly look at the summary statistics for some of the variables in the data. Table 1 shows some summary statistics for the parent data (data of the 49 mothers only). We can see that the mean age of the mothers is 38, and most of the mothers are fully employed (0 = not employed: 29%, 1 = part-time: 17%, 2 = full-time: 54%). The `mom_numcig` variable is the mothers' number of cigarettes per day, and we see that 67% of the mothers self-reported zero cigarette per day and very few that have 8 or more cigarettes per day. The `income` variable represents the family's estimated annual household income and the mean income for this data set is 57947 with a large standard deviation of 51607. From the employment status, 29% of the mothers are not employed and 17% have part-time jobs, and this could be the reason for the high variance in income. The `mom_smoke` variables are the self-reported smoking status at and after pregnancy. It can be seem that 25% to 31% of the mothers are smokers during pregnancy, and 24% to 40% of the mothers are smokers after pregnancy. We do see that the percentages of mother smokers are slightly higher after pregnancy, for instance `mom_smoke_pp6mo` is 40% which means 40% of the mothers are smoking at 6 months postpartum while the highest proportion during preganancy is 31%. The `cotimean` variables are urine cotinine (nicotine metabolite) during and after pregnancy. In this case, higher cotinine levels mean heavier smokers or heavier smoke exposures. We can see that the mean for `cotimean_pp6mo` is 100. This is a similar pattern we see in the `mom_smoke` variables that the smoking percentage tends to be lower during pregnancy and in the `cotimean` variables, the cotinine level tends to be lower during pregnancy. However, since some of the data relating to SDP are self-reported there could be bias in the data that affects the overall data quality.

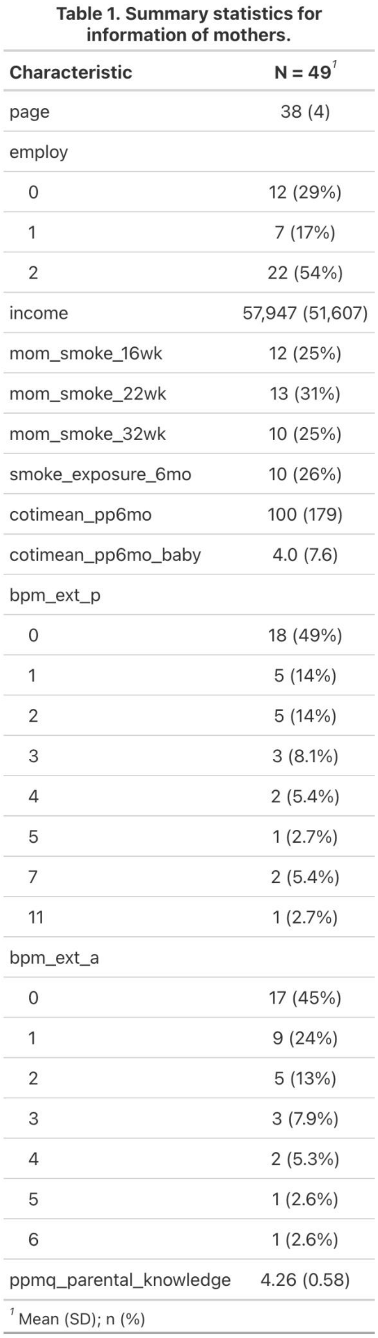


Table 1. Summary statistics for information of mothers.

Characteristic	N = 49[1]
page	38 (4)
employ	
0	12 (29%)
1	7 (17%)
2	22 (54%)
income	57,947 (51,607)
mom_smoke_16wk	12 (25%)
mom_smoke_22wk	13 (31%)
mom_smoke_32wk	10 (25%)
smoke_exposure_6mo	10 (26%)
cotimean_pp6mo	100 (179)
cotimean_pp6mo_baby	4.0 (7.6)
bpm_ext_p	
0	18 (49%)
1	5 (14%)
2	5 (14%)
3	3 (8.1%)
4	2 (5.4%)
5	1 (2.7%)
7	2 (5.4%)
11	1 (2.7%)
bpm_ext_a	
0	17 (45%)
1	9 (24%)
2	5 (13%)
3	3 (7.9%)
4	2 (5.3%)
5	1 (2.6%)
6	1 (2.6%)
ppmq_parental_knowledge	4.26 (0.58)

[1] Mean (SD); n (%)

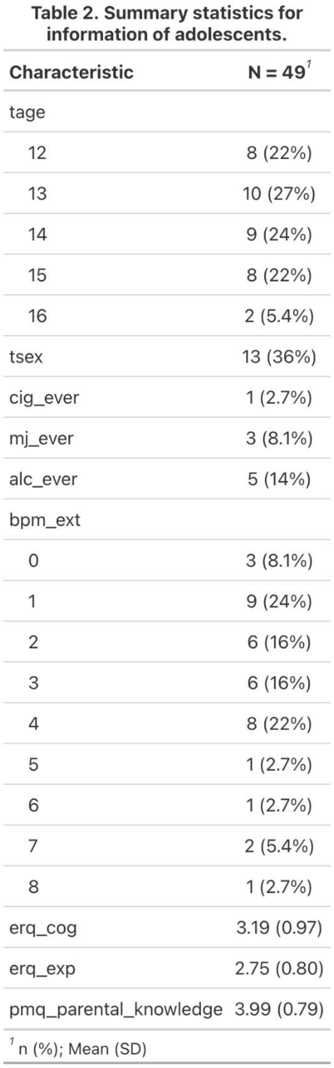


Table 2. Summary statistics for information of adolescents.

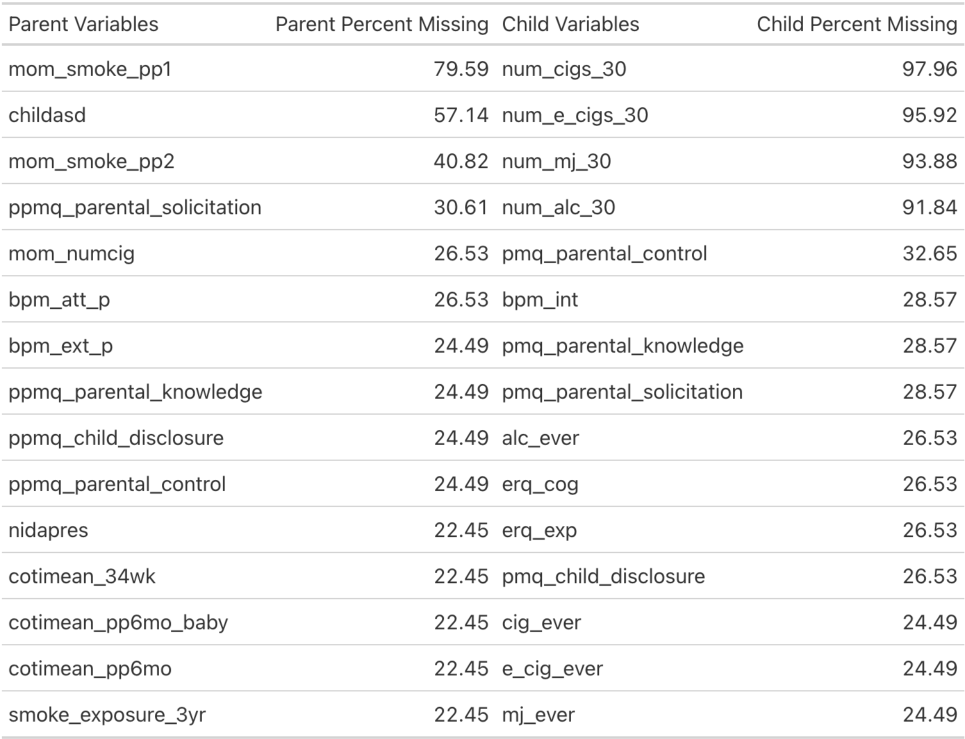
Characteristic	N = 49[1]
tage	
12	8 (22%)
13	10 (27%)
14	9 (24%)
15	8 (22%)
16	2 (5.4%)
tsex	13 (36%)
cig_ever	1 (2.7%)
mj_ever	3 (8.1%)
alc_ever	5 (14%)
bpm_ext	
0	3 (8.1%)
1	9 (24%)
2	6 (16%)
3	6 (16%)
4	8 (22%)
5	1 (2.7%)
6	1 (2.7%)
7	2 (5.4%)
8	1 (2.7%)
erq_cog	3.19 (0.97)
erq_exp	2.75 (0.80)
pmq_parental_knowledge	3.99 (0.79)

[1] n (%); Mean (SD)

Now looking at Table 2, we have information of the selected variables of the child data (data of the 49 adolescents only). We can see that most of the child are white (39%) and black (31%). The `cig_ever` variable is whether the child ever tried cigarette smoking, and only 1 child have smoking experience and 12 data points missing. The `e_cig_ever` variable is whether the child ever tried e-cigarette smoking, and 3 child have e-cigarrete smoking experience. The `mj_ever` variable is whether the child ever tried marijuana, and also 3 child responded yes. The `alc_ever` variable is whether the child ever tried alcohol, and 5 child responded yes. Even though from Table 1, we see that the mom smoking percentages at pregnancy are between 25% to 31%, the child's smoking, substance use, and alcohol use percentages are relatively low.

Missing Data

Table 3. Percent missing for all variables in the parent and child data.



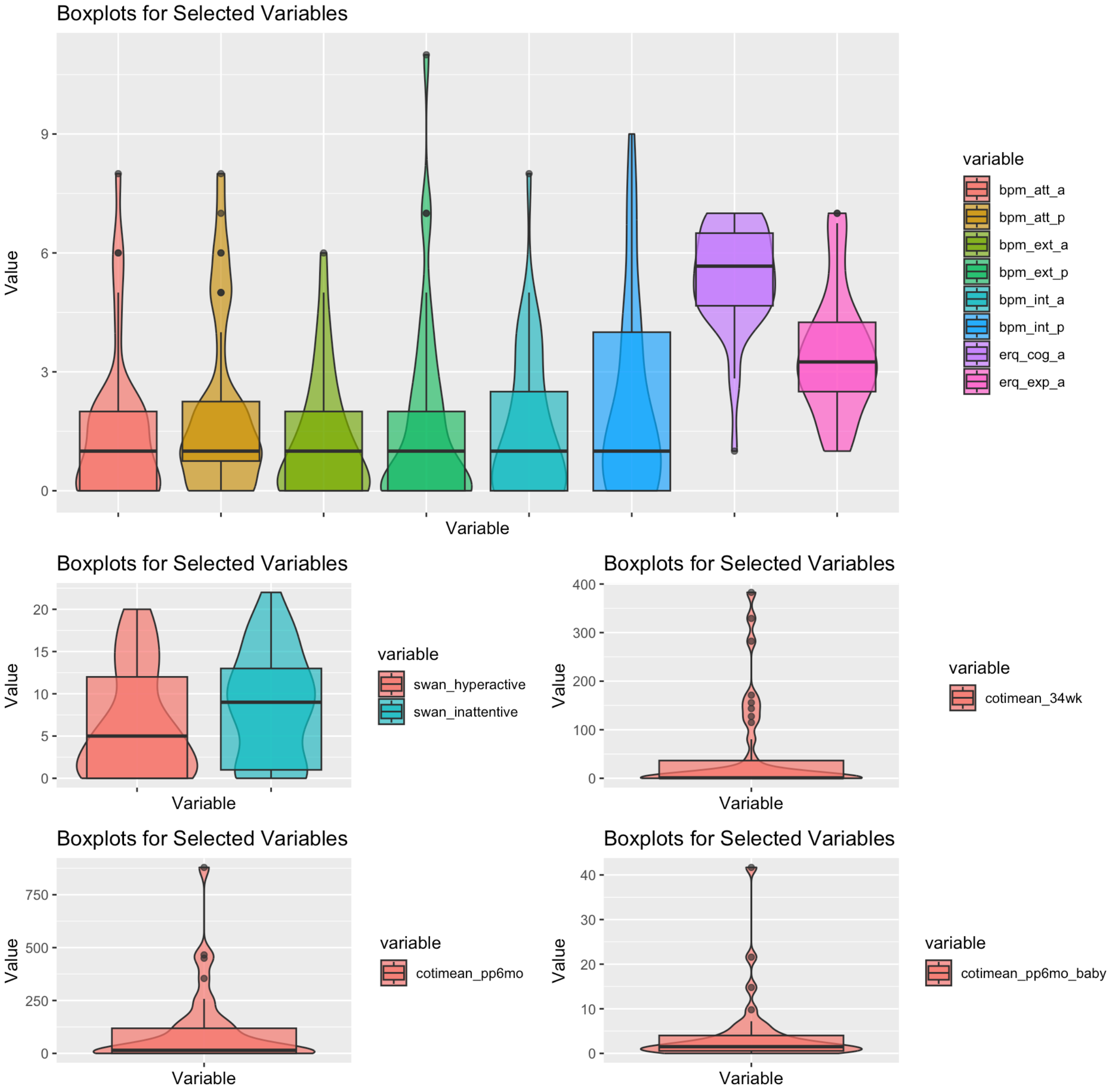
Parent Variables	Parent Percent Missing	Child Variables	Child Percent Missing
mom_smoke_pp1	79.59	num_cigs_30	97.96
childasd	57.14	num_e_cigs_30	95.92
mom_smoke_pp2	40.82	num_mj_30	93.88
ppmq_parental_solicitation	30.61	num_alc_30	91.84
mom_numcig	26.53	pmq_parental_control	32.65
bpm_att_p	26.53	bpm_int	28.57
bpm_ext_p	24.49	pmq_parental_knowledge	28.57
ppmq_parental_knowledge	24.49	pmq_parental_solicitation	28.57
ppmq_child_disclosure	24.49	alc_ever	26.53
ppmq_parental_control	24.49	erq_cog	26.53
nidapres	22.45	erq_exp	26.53
cotimean_34wk	22.45	pmq_child_disclosure	26.53
cotimean_pp6mo_baby	22.45	cig_ever	24.49
cotimean_pp6mo	22.45	e_cig_ever	24.49
smoke_exposure_3yr	22.45	mj_ever	24.49

Since we are interested in the effects of SDP/ETS, it is important to examine whether we have sufficient data to support any hypotheses. We can first explore the missing patterns in the data. From Table 3, we have the top 15 percent missing variables included. For the list of parent variables, we see that the variable with highest missing percentage (79.59%) is `mom_smoke_pp1`. For the list of child variables, we see that the variables with highest missing percentages are `num_cigs_30` (97.96%), `num_e_cigs_30` (95.92%), `num_mj_30` (93.88%), and `num_alc_30` (91.84%), while other variables are mostly ranging from 24% to 32% of missingness. The high percentages here is actually caused by the variables `cig_ever`, `e_cig_ever`, `mj_ever`, and `alc_ever` since most of the adolescents responded "No" and were not asked to answer the number of cigarette, marijuana, or alcohol used in the past 30 days. Even though the missing percentages of other variables in Table 3 are mostly ranging between about 20% to 30%, we might have some difficulties in finding the effects and interrelatedness of SDP/ETS, self-regulation, externalizing behavior, and substance use. We only have a total of 49 pairs of mother and adolescent, and on top of that, not many of the adolescents use cigarettes, marijuana, or alcohol, and most mothers are in non-smoking status (self-reported) during the period of the study. Given the case, we will have to focus on other variables such as the cotimeans, ppmq, bpm, and erq to further examine the effects and interrelatedness.

For identifying the missing mechanism, we can not make the MCAR assumption since it is hardly realistic to assume the missingness is unrelated of any unobserved data. However, it is also difficult to distinguish the missing mechanism between MAR and MNAR for our data set because we have multivariable missingness. For instance, it is unclear whether the high missingness in `mom_smoke_pp1` is directly related to pregnancy status (pregnant or postpartum) or other indirect observed variables. But if we assume MNAR, then the distribution of the missing observations do not only depend on the observed values but also the unobserved values. Thus, for the purpose of this project, we can continue with the assumption that missing data are unrelated to unobserved values given the observed data (MAR).

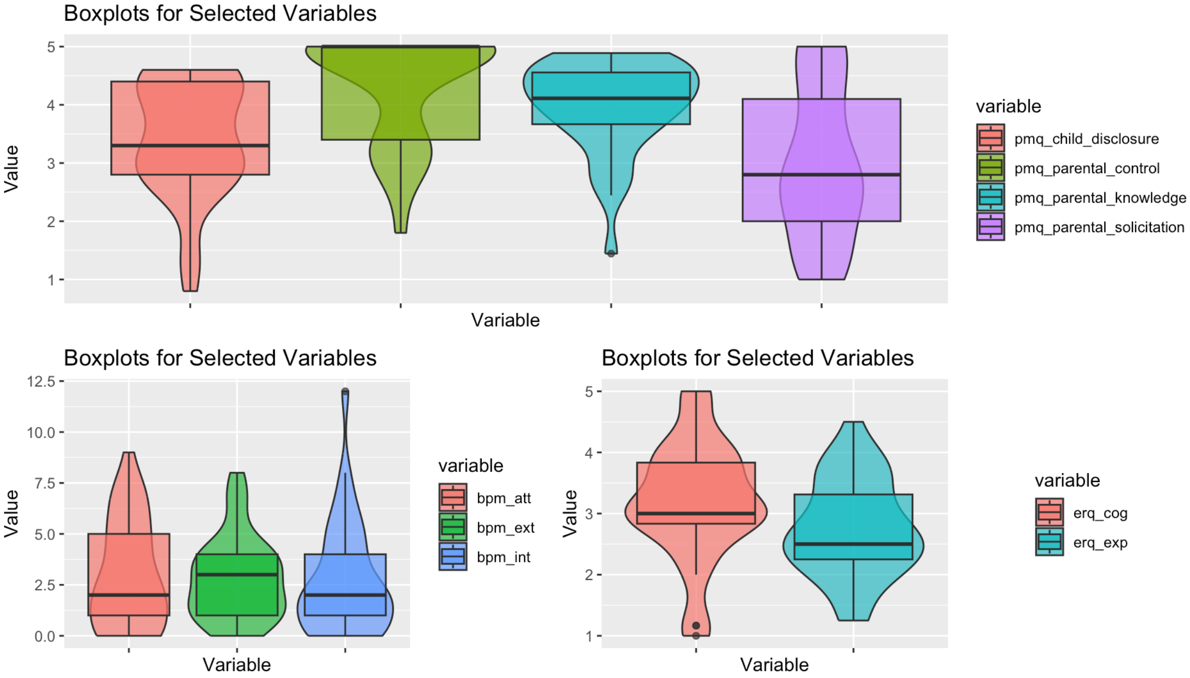
After exploring the missing patterns, we can now take a closer look at the variables of interest (smoke exposures, self-regulations, etc) by creating boxplots. Figure 1 shows the distributions and boxplots of the selected variables in parent data. We can see that the brief problem monitor variables, most of the responses are pretty low and the distributions are right skewed. And we also see some outliers especially in `bpm_ext_a`, the value is 11 which is very high. But overall, the bpm values are low meaning most of the mothers or adolescents are doing well in terms of problems related to attention, externalizing, and internalizing. For the cotimeans variables, there are some very high outliers in both cotinine after 6 month postpartum and cotinine at 34 weeks of gestation. Even though the distributions for the cotimeans are still right skewed, there are many high outliers which imply high smoke exposures or smoking duration of the child and mother. These high cotinine levels might also be the cause of the high outliers in the brief problem monitor reponses.

Figure 1: Multiple box plots for selected variables in parent data.



Now we can take a closer look at the selected variables in child data. We can immediately see that there are fewer outliers compared to boxplots in the parent data. But from the boxplots of brief problem monitor variables, there is a very high value in `bpm_int` . And we some see differences that the distributions are less right skewed and mostly centered at a higher mean value compared to the bpm boxplots in parent data. For the parental knowledge variables, we see that they are mostly centered at higher values (greater than 2.5 out of 5) which means most of the adolescents are having proper parental care or good relationships with their parents. For the 2 emotion regulation variables, the distributions are centered at an average value meaning average control over emotions in the adolescents. We can even see some very low outliers in `erq_cog`, which means very poor control over emotions. From these boxplots, we can see that high smoke exposures from mothers might affect child's ability to handle self-regulation problems since we do observe some very poor values given low mother smoking duration (24% - 29% smoking during pregnancy and 40% 6 months after postpartum) seen from Table 1.

Figure 2: Multiple box plots for selected variables in child data.



Correlations and Interrelatedness

After having a solid understanding of the variables in the data set, we can now examine the interrelations among the variables of interest. First, we can use the smoke exposures variables to create two new variables called SDP Intensity and ETS Intensity to investigate whether higher smoke exposures duration causes poorer ability to handle self-regulation problems for adolescents. SDP and ETS Intensities are created by filtering the parent data that include no missing information in the variables: `mom_smoke_16wk`, `mom_smoke_22wk`, `mom_smoke_32wk`, `smoke_exposure_6mo`, `smoke_exposure_12mo`, `smoke_exposure_2yr`, `smoke_exposure_3yr`, `smoke_exposure_4yr`, and `smoke_exposure_5yr`. SDP intensity is created by using the `mom_smoke` variables and ranges from 0 to 4, where 4 means the highest smoke during pregnancy intensity. ETS intensity is created by using the `smoke_exposure` variables and ranges from 0 to 6, where 6 means the highest environmental tobacco smoke intensity. Figure 3 is a correlation plot for interrelatedness between prenatal and postnatal smoke exposure. We can see that the correlation (R = 0.55) is moderately strong and it is significant with a p-value less than 0.05 alpha level. This figure suggests that ETS and SDP are interrelated.

Figure 3: Correlation between SDP and ETS Intensities

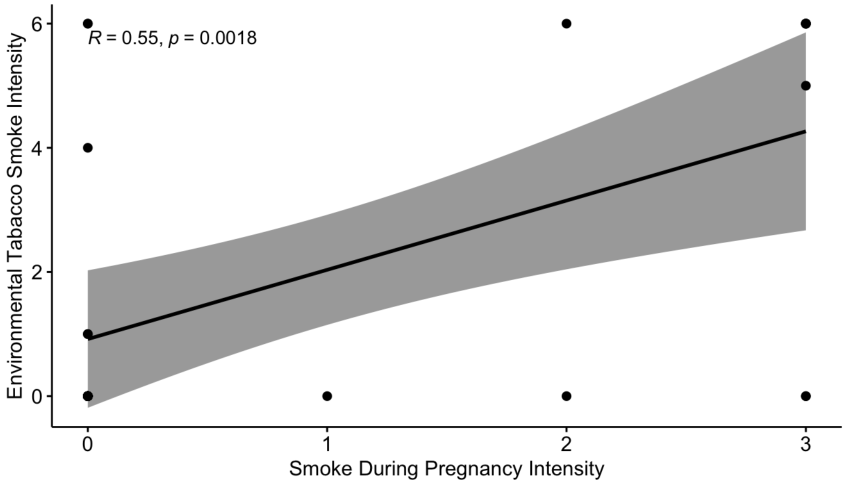
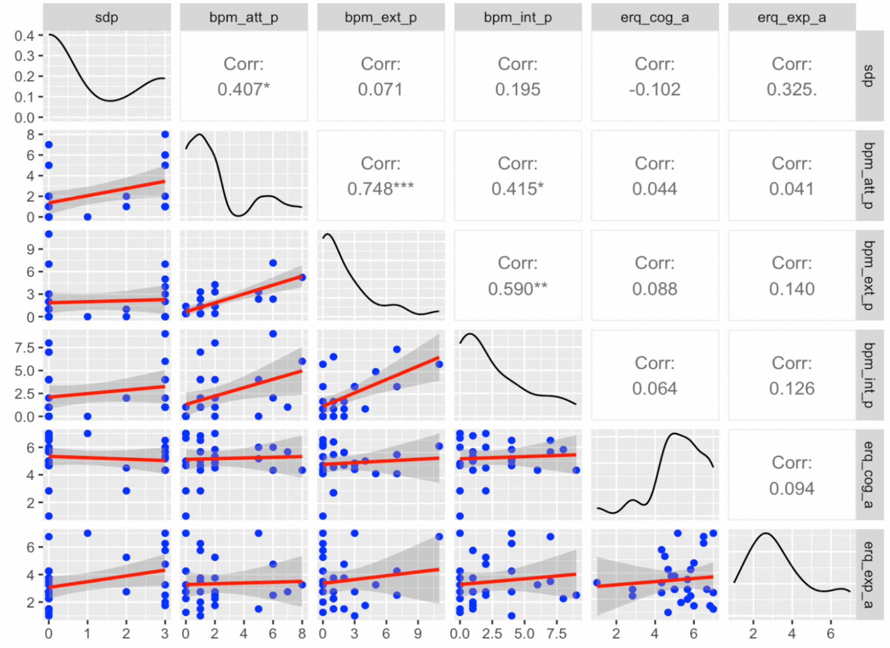


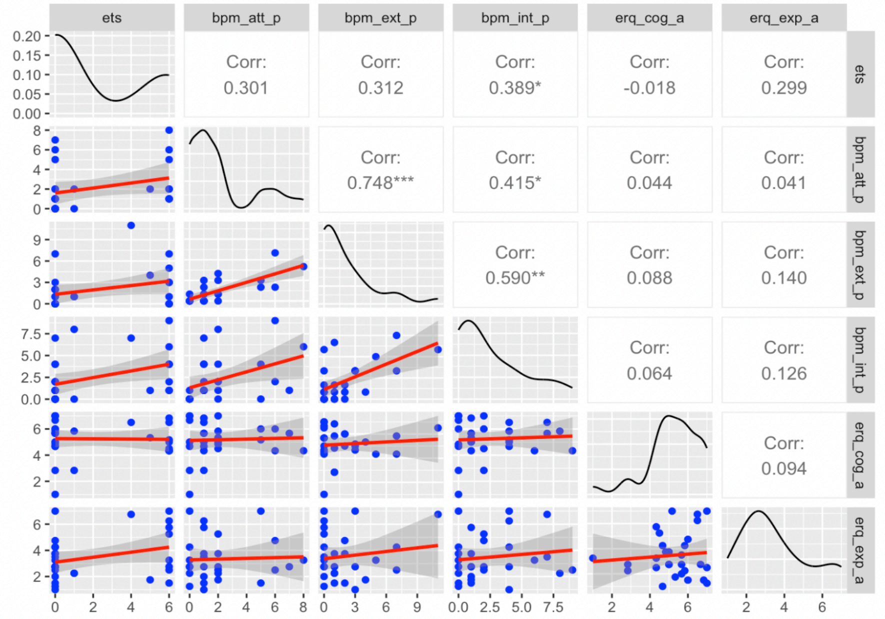
Figure 4 shows the scatter plots of SDP intensity and self-regulation and externalizing behavior variables along with their correlation values on the upper side. In general, all the self-regulation variables are positively correlated with SDP. The highest correlation for SDP intensity is 0. 407 with `bpm_att_p`. This suggests a higher SDP intensity is positively correlated with a higher scores in attention problems on child. The other strong and the highest correlation in Figure 3 is 0.748 which is between `bpm_ext_p` and `bpm_att_p`. This implies that a higher value in attention problems is strongly correlated with a higher value in externalizing problems positively. Even though other variables are also positively correlated with SDP intensity, they are pretty weak correlations which could be due to the high number of 0 smoke exposures duration.

Figure 4: Scatter plots and correlations for SDP Intensity and self-regulation and externalizing behavior variables



In addition to SDP intensity, we can also examine the ETS intensity. From Figure 5, we can observe that all correlations are also positive. But there are not strong positive correlations between ETS intensity and self-regulation and externalizing behavior variables. `bpm_att_p` (R = 0.301) and `bpm_int_p` (R = 0.389) are somewhat correlated with it. From the distribution of ETS intensity, we can see that it is heavily right skewed with most of the mothers self-report no smoke exposure. This makes the interpretation of the correlations difficult since we do not have a lot of smoke exposure cases or smoking during pregnancy cases in the higher end (32 weeks of smoking during pregnancy or 22 weeks of smoking during pregnancy). Even though we do not observe very strong correlations in both Figure 5 and Figure 4, correlation does not directly means causation so further analyses are needed, and regressions might be more helpful in seeing the actual main effects of SDP/ETS.

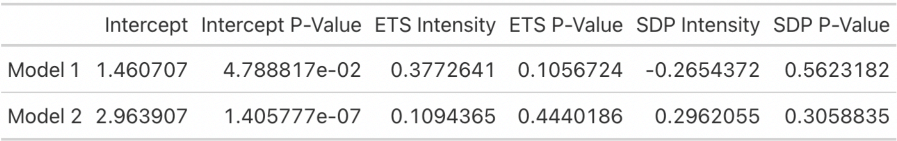
Figure 5: Scatter plots and correlations for ETS Intensity and self-regulation and externalizing behavior variables



Regressions

Based on the correlations we see from Figure 5 and Figure 4, we can also construct simple linear regression model to see whether there is significant main effects of SDP/ETS on self-regulation and externalizing behavior variables. Figure 6 shows the linear regression results for Model 1 and Model 2. Both Model 1 and Model 2 included only the ETS intensity and SDP intensity as predictors so that main effects can be compared between the two intensities. The dependent variable in Model 1 is `bpm_ext_p` and the dependent variable in Model 2 is `erq_exp_a`.

Table 6. Regression coefficients for simple linear Model 1 and Model 2.



	Intercept	Intercept P-Value	ETS Intensity	ETS P-Value	SDP Intensity	SDP P-Value
Model 1	1.460707	4.788817e-02	0.3772641	0.1056724	-0.2654372	0.5623182
Model 2	2.963907	1.405777e-07	0.1094365	0.4440186	0.2962055	0.3058835

From Model 1 we can see that there are actually no significant main effects from the independent variables. This is expected since we do not have enough data and lacked sufficient cases for SDP and ETS. The coefficient estimate for ETS intensity has the lowest p-value of 0.10 and has the largest positive effect on externalizing behavior. The results for Model 2 is also shown. Similar to Model 1, no coefficient estimates are significant. The two fitted linear regression models above are just for exploratory purposes, to get more specific or accurate results, we need to consider more types of regressions and more careful selection for independent predictors.

Conclusion

We are able to see some patterns of the effects of SDP/ETS that higher smoke exposure duration and smoking during pregnancy might lead to poorer ability of adolescents to handle self-regulation problems. For instance, higher smoke exposure duration is positively correlates with higher expressive suppression. Higher expressive suppression is also positively correlates with higher values for internalizing problems. However, the limitation of our data is that the proportion of smoking mothers and non-smoking mothers is not balanced with most of the mothers not smoking during and after pregnancy. This makes investigation of the effects of SDP/ETS very challenging since the mother smoking cases are almost considered as outliers. Since we only have a total 98 people in the data, it will also be beneficial if we can include more pairs of mothers and adolescents so that we have sufficient data to further analyze the effects and interrelatedness among the variables. Self-reporting data is also another limitation of this project since the data were collected retrospectively, and it would be better if more accurate information similar to the urine cotinine can be obtained in the future studies. Having more reliable measures relating to the topic of interest is important for further statistical analyses.

Reference

Centers for Disease Control and Prevention. (2020, April 28). *Smoking during pregnancy*. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/basic_information/health_effects/pregnancy/index.html

Micalizzi, L. (2023, Oct 16). *Prenatal Tobacco Exposure & Child Outcomes* [PowerPoint slides]. School of Public Health, Brown University.

Appendix

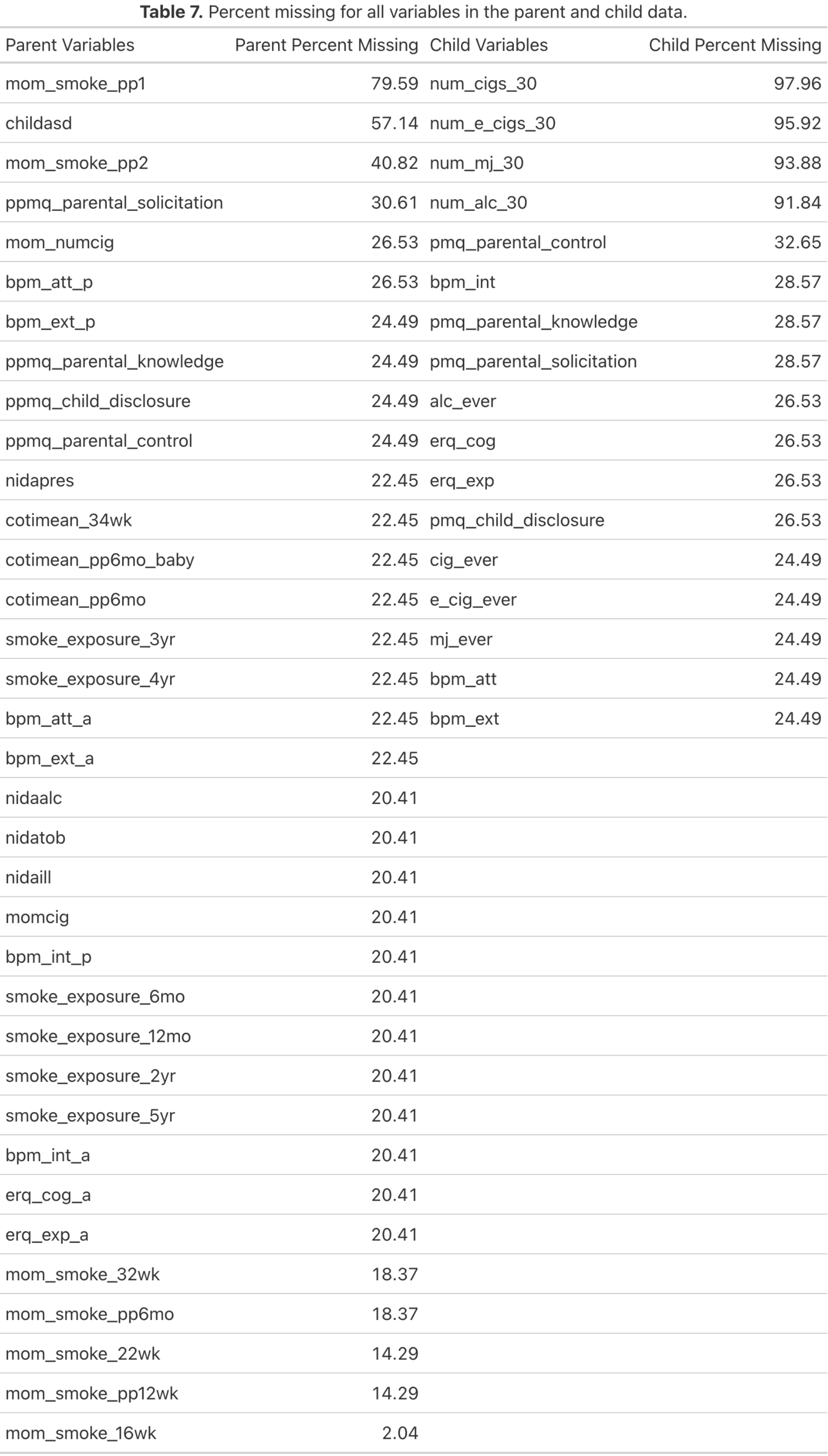


Table 7. Percent missing for all variables in the parent and child data.

Parent Variables	Parent Percent Missing	Child Variables	Child Percent Missing
mom_smoke_pp1	79.59	num_cigs_30	97.96
childasd	57.14	num_e_cigs_30	95.92
mom_smoke_pp2	40.82	num_mj_30	93.88
ppmq_parental_solicitation	30.61	num_alc_30	91.84
mom_numcig	26.53	pmq_parental_control	32.65
bpm_att_p	26.53	bpm_int	28.57
bpm_ext_p	24.49	pmq_parental_knowledge	28.57
ppmq_parental_knowledge	24.49	pmq_parental_solicitation	28.57
ppmq_child_disclosure	24.49	alc_ever	26.53
ppmq_parental_control	24.49	erq_cog	26.53
nidapres	22.45	erq_exp	26.53
cotimean_34wk	22.45	pmq_child_disclosure	26.53
cotimean_pp6mo_baby	22.45	cig_ever	24.49
cotimean_pp6mo	22.45	e_cig_ever	24.49
smoke_exposure_3yr	22.45	mj_ever	24.49
smoke_exposure_4yr	22.45	bpm_att	24.49
bpm_att_a	22.45	bpm_ext	24.49
bpm_ext_a	22.45		
nidaalc	20.41		
nidatob	20.41		
nidaill	20.41		
momcig	20.41		
bpm_int_p	20.41		
smoke_exposure_6mo	20.41		
smoke_exposure_12mo	20.41		
smoke_exposure_2yr	20.41		
smoke_exposure_5yr	20.41		
bpm_int_a	20.41		
erq_cog_a	20.41		
erq_exp_a	20.41		
mom_smoke_32wk	18.37		
mom_smoke_pp6mo	18.37		
mom_smoke_22wk	14.29		
mom_smoke_pp12wk	14.29		
mom_smoke_16wk	2.04		

Code Appendix

```
# libraries
library(tidyverse)
library(data.table)
library(GGally)
library(ggplot2)
library(gtsummary)
library(tableone)
library(cowplot)
library(magick)
library(patchwork)
library(gridExtra)

# read data
data <- read.csv("~/Downloads/project1.csv")

# dim(data)
# head(data)

# looking at mom_numcig
# in order to look at correlations, convert this variable to numeric
data$mom_numcig[data$mom_numcig %in% c("" , "None" , "44989" ,
                                       "2 black and miles a day")] <- NA
data$mom_numcig[data$mom_numcig == "20-25"] <- 20
data$mom_numcig <- as.numeric(data$mom_numcig)

# calculate missing percentage for each variables
missing <- round(apply(data, 2, function(x) sum(is.na(x)))/nrow(data), 4)
missing <- missing * 100

# remove row with all missing data
na_ind <- apply(data, 1, function(x) all(is.na(x)))
data <- data[!na_ind, ]
# missing

# create parent data and adjust mom smoke variables
parent_data <- data[ ,1:51] %>%
  mutate(mom_smoke_16wk = case_when(mom_smoke_16wk == "2=No" ~ 0,
                                    mom_smoke_16wk == "1=Yes" ~ 1,
                                    mom_smoke_16wk == "" ~ NA)) %>%
  mutate(mom_smoke_22wk = case_when(mom_smoke_22wk == "2=No" ~ 0,
                                    mom_smoke_22wk == "1=Yes" ~ 1,
                                    mom_smoke_22wk == "" ~ NA)) %>%
  mutate(mom_smoke_32wk = case_when(mom_smoke_32wk == "2=No" ~ 0,
                                    mom_smoke_32wk == "1=Yes" ~ 1,
                                    mom_smoke_32wk == "" ~ NA)) %>%
  mutate(mom_smoke_pp1 = case_when(mom_smoke_pp1 == "2=No" ~ 0,
                                    mom_smoke_pp1 == "1=Yes" ~ 1,
                                    mom_smoke_pp1 == "" ~ NA)) %>%
  mutate(mom_smoke_pp2 = case_when(mom_smoke_pp2 == "2=No" ~ 0,
                                    mom_smoke_pp2 == "1=Yes" ~ 1,
                                    mom_smoke_pp2 == "" ~ NA)) %>%
  mutate(mom_smoke_pp12wk = case_when(mom_smoke_pp12wk == "2=No" ~ 0,
                                    mom_smoke_pp12wk == "1=Yes" ~ 1,
                                    mom_smoke_pp12wk == "" ~ NA)) %>%
  mutate(mom_smoke_pp6mo = case_when(mom_smoke_pp6mo == "2=No" ~ 0,
                                    mom_smoke_pp6mo == "1=Yes" ~ 1,
                                    mom_smoke_pp6mo == "" ~ NA))

# convert income from character type to numeric
parent_data$income <- as.numeric(parent_data$income)
child_data <- data[ ,c(1, 52:78)]

# dim(parent_data)
# dim(child_data)
#
# head(parent_data)

# long data for continuous variables
parent_long <- parent_data %>%
  select(parent_id, page, income, momcig, mom_numcig, cotimean_34wk,
         cotimean_pp6mo_baby, cotimean_pp6mo, swan_inattentive,
         swan_hyperactive, bpm_att_p, bpm_ext_p, bpm_int_p,
         ppmq_parental_knowledge, ppmq_child_disclosure, ppmq_parental_solicitation,
         ppmq_parental_control, bpm_att_a, bpm_ext_a, bpm_int_a, erq_cog_a,
         erq_exp_a)
parent_long <- pivot_longer(parent_long, cols = c(2:22), names_to = "variable",
                            values_to = "value")

child_long <- child_data %>%
  select(parent_id, bpm_att, bpm_ext, bpm_int, erq_cog, erq_exp, pmq_parental_knowledge,
         pmq_child_disclosure, pmq_parental_solicitation, pmq_parental_control)
child_long <- pivot_longer(child_long, cols = c(2:10), names_to = "variable",
                            values_to = "value")

# calculate missing percentage for each variables
parent_missing <- round(apply(parent_data, 2, function(x) sum(is.na(x)))/nrow(parent_data)
parent_missing <- parent_missing * 100

parent_missing_table <- as.data.frame(parent_missing)
parent_missing_table$variable <- colnames(parent_data)
parent_missing_table <- parent_missing_table[ c(15:31, 34:51),] %>%
  relocate(parent_missing, .after = variable) %>%
  select(parent_missing)

#kableone(parent_missing_table)

# calculate missing percentage for each variables
child_missing <- round(apply(child_data, 2, function(x) sum(is.na(x)))/nrow(child_data), 4
child_missing <- child_missing * 100

child_missing_table <- as.data.frame(child_missing)
child_missing_table$variable <- colnames(child_data)
child_missing_table <- child_missing_table[ c(12:28),] %>%
  relocate(child_missing, .after = variable)

# bind missing percentages to create table
child_na <- data.frame(variable = rep("", 18),
                          child_missing = rep("", 18))
child_missing_table <- rbind(child_missing_table, child_na)

parent_child_missing <- cbind(parent_missing_table, child_missing_table)
#kableone(child_missing_table)

# descriptive summary for selected variables of mothers
table1 <- parent_data %>%
  tbl_summary(include = c(page, paian, pasian, pnhpi,
                          pblack, pwhite, prace_other, mom_numcig,
                          mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk,
                          mom_smoke_pp1, mom_smoke_pp2, mom_smoke_pp12wk,
                          mom_smoke_pp6mo, cotimean_34wk, cotimean_pp6mo,
                          cotimean_pp6mo_baby),
              statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n}  ({p}%)"
    ),
    missing_text = "NA")%>%
  modify_caption("**Table 1. Summary statistics for selected variables/information of moth
table1

gt::gtsave(as_gt(table1), file ="project1_table1.png")

# create boxplots for continuous variables
parent_long1 <- parent_long[parent_long$variable %in% c("bpm_att_p",
                                                      "bpm_ext_p", "bpm_int_p",
                                                      "bpm_att_a", "bpm_ext_a",
                                                      "bpm_int_a", "erq_cog_a",
                                                      "erq_exp_a"),]
parent_long2 <- parent_long[parent_long$variable %in% c("ppmq_parental_knowledge",
                                                        "ppmq_child_disclosure",
                                                        "ppmq_parental_solicitation",
                                                        "ppmq_parental_control"),]
parent_long3 <- parent_long[parent_long$variable %in% c("swan_inattentive",
                                                        "swan_hyperactive"),]
parent_long4 <- parent_long[parent_long$variable == "cotimean_34wk",]
parent_long5 <- parent_long[parent_long$variable == "cotimean_pp6mo",]
parent_long6 <- parent_long[parent_long$variable == "cotimean_pp6mo_baby",]
# a function to create boxplots
boxplot_func <- function(df) {
#' @description creat multiple boxplots for the given data
#' @param df a dataframe
#' @return a set of boxplots for the given data

  ggplot(df, aes(variable, value, fill = variable)) +
    geom_violin(alpha = 0.7) + # visualizes the shape of the distribution as well
    geom_boxplot(alpha = 0.7) +
    scale_x_discrete(name = "Variable") +
    scale_y_continuous(name = "Value") +
    ggtitle("Boxplots for Selected Variables") +
    theme(axis.text.x = element_blank())
}

parent_boxplot1 <- boxplot_func(parent_long1)
```

```
parent_boxplot2 <- boxplot_func(parent_long2)
parent_boxplot3 <- boxplot_func(parent_long3)
parent_boxplot4 <- boxplot_func(parent_long4)
parent_boxplot5 <- boxplot_func(parent_long5)
parent_boxplot6 <- boxplot_func(parent_long6)
# parent_boxplot1
# parent_boxplot2
# parent_boxplot3
# parent_boxplot4
# parent_boxplot5
# parent_boxplot6

project1_boxplots1 <- parent_boxplot1 / (parent_boxplot3 + parent_boxplot4 +
                                           parent_boxplot5 +parent_boxplot6) +
  plot_annotation( title = 'Figure 1: Multiple box plots for selected variables in parent
                   theme = theme(plot.title = element_text(size = 13)))
project1_boxplots1
ggsave("project1_boxplots1.jpg", project1_boxplots1)

table2 <- child_data %>%
  tbl_summary(include = c(taian, tasian, tnhpi,
                          tblack, twhite, trace_other, cig_ever, e_cig_ever,
                          mj_ever, alc_ever),
              statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n}  ({p}%)"
    ),
    missing_text = "NA")%>%
  modify_caption("**Table 2. Summary statistics for selected variables/information of adol
table2

gt::gtsave(as_gt(table2), file ="project1_table2.png")

project1_table1 <- ggdraw() + draw_image("project1_table1.png")
project1_table2 <- ggdraw() + draw_image("project1_table2.png", scale = 0.8)
project1_table12 <- plot_grid(project1_table1, project1_table2)
ggsave("project1_table12.jpg", project1_table12)

# create boxplots for continuous variables
child_long1 <- child_long[child_long$variable %in% c("bpm_att",
                                                      "bpm_ext", "bpm_int"),]
```

```
child_long2 <- child_long[child_long$variable %in% c("erq_cog", "erq_exp"),]
child_long3 <- child_long[child_long$variable %in% c("pmq_parental_knowledge",
                                                        "pmq_child_disclosure",
                                                        "pmq_parental_solicitation",
                                                        "pmq_parental_control"),]

child_boxplot1 <- boxplot_func(child_long1)
child_boxplot2 <- boxplot_func(child_long2)
child_boxplot3 <- boxplot_func(child_long3)
child_boxplot1
child_boxplot2
child_boxplot3

project1_boxplots2 <- child_boxplot3 / (child_boxplot1 + child_boxplot2) +
  plot_annotation( title = 'Figure 2: Multiple box plots for selected variables in child d
project1_boxplots2
ggsave("project1_boxplots2.jpg", project1_boxplots2)


# check general smoke exposure child
smoke_exposure_df <- data[ , c("parent_id", "smoke_exposure_6mo", "smoke_exposure_12mo",
                          "smoke_exposure_2yr", "smoke_exposure_3yr",
                          "smoke_exposure_4yr", "smoke_exposure_5yr")]

smoke_exposure_df <- melt(smoke_exposure_df, id.vars = "parent_id",
                          variable.name = "smoke_exposure_time")
smoke_exposure_df <- smoke_exposure_df %>% group_by(smoke_exposure_time) %>%
  summarize(yes = length(value[value==1]), no = length(value[value ==0]))

smoke_exposure_df <- melt(smoke_exposure_df, id.vars = "smoke_exposure_time",
                          variable.name = "smoke_exposure_bin")

ggplot(smoke_exposure_df, aes(x = smoke_exposure_time,
                 y = value, fill = smoke_exposure_bin)) +
  geom_bar(position="dodge", stat = "identity")

# data <- data %>%
#   mutate(se_bin = case_when(smoke_exposure_6mo == 1 ~ "se_6mo_yes",
#                    smoke_exposure_6mo == 0 ~ "se_6mo_no",
#                    smoke_exposure_12mo == 1 ~ "se_12mo_yes",
#                    smoke_exposure_12mo == 0 ~ "se_12mo_no",
#                    smoke_exposure_2yr == 1 ~ "se_2yr_yes",
#                    smoke_exposure_2yr == 0 ~ "se_2yr_no",
#                    smoke_exposure_3yr == 1 ~ "se_3yr_yes",
#                    smoke_exposure_3yr == 0 ~ "se_3yr_no",
#                    smoke_exposure_4yr == 1 ~ "se_4yr_yes",
#                    smoke_exposure_4yr == 0 ~ "se_4yr_no",
#                    smoke_exposure_5yr == 1 ~ "se_5yr_yes",
#                    smoke_exposure_5yr == 0 ~ "se_5yr_no",
#                    is.na(smoke_exposure_6mo) | is.na(smoke_exposure_12mo) |
#                    is.na(smoke_exposure_2yr) | is.na(smoke_exposure_3yr) |
#                    is.na(smoke_exposure_4yr) | is.na(smoke_exposure_5yr) ~ NA))

# data <- data %>%
#   mutate(se_ordinal = case_when(
#                    smoke_exposure_5yr == 1 ~ 6,
#                    smoke_exposure_4yr == 1 ~ 5,
#                    smoke_exposure_3yr == 1 ~ 4,
#                    smoke_exposure_2yr == 1 ~ 3,
#                    smoke_exposure_12mo == 1 ~ 2,
#                    smoke_exposure_6mo == 1 ~ 1,
#                    # smoke_exposure_5yr == 0 ~ 0,
#                    # smoke_exposure_4yr == 0 ~ 0,
#                    # smoke_exposure_3yr == 0 ~ 0,
#                    # smoke_exposure_2yr == 0 ~ 0,
#                    # smoke_exposure_12mo == 0 ~ 0,
#                    smoke_exposure_6mo == 0 ~ 0,
#                    is.na(smoke_exposure_6mo) | is.na(smoke_exposure_12mo) |
#                    is.na(smoke_exposure_2yr) | is.na(smoke_exposure_3yr) |
#                    is.na(smoke_exposure_4yr) | is.na(smoke_exposure_5yr) ~ NA))

# calculate smoke exposure duration
data$smoke_exposure_duration <- (data$smoke_exposure_6mo - 0.5) * (data$smoke_exposure_6mo
  (data$smoke_exposure_12mo - 0.5) * (data$smoke_exposure_12mo != 0) +
  data$smoke_exposure_2yr + data$smoke_exposure_3yr + data$smoke_exposure_4yr +
  data$smoke_exposure_5yr

# calculate mom smoke duration during pregnancy
data$mom_smoke_pregnant <- parent_data$mom_smoke_16wk +
  parent_data$mom_smoke_22wk + parent_data$mom_smoke_32wk

# calculate mom smoke duration after pregnancy
```

```r
data$mom_smoke_post <- parent_data$mom_smoke_pp12wk + parent_data$mom_smoke_pp6mo

# a function to customize fitted lines in scatter plots
lower_func <- function(data, mapping, method = "lm", ...) {
#' @description customize fitted lines in ggally plots
#' @param data input data for plotting
#' @param mapping the mapping for variables in the input data
#' @param method the method to be used for fitted lines
#' @return a ggplot with fitted lines and the corresponding mapping

  p <- ggplot(data = data, mapping = mapping) +
    geom_point(colour = "blue") +
    geom_smooth(method = method, color = "red", ...)

  return(p)
}

# use ggally for correlations and scatterplot
ggpair_df3 <- data[, c("smoke_exposure_duration", "bpm_att", "bpm_ext",
                          "bpm_int", "erq_cog", "erq_exp")]

corplot1 <- ggpairs(ggpair_df3, columns = 1:ncol(ggpair_df3),
        lower = list(continuous = wrap(lower_func, method = "lm")),
        title = "Figure 3: Scatter plots and correlations for smoke exposures duration and
        axisLabels = "show", columnLabels = colnames(ggpair_df3))
corplot1 <- corplot1 + theme(plot.title = element_text(size = 10))
ggsave("corplot1.jpg", corplot1)

# use ggally for correlations and scatterplot
ggpair_df4 <- data[, c("mom_smoke_pregnant", "bpm_att", "bpm_ext",
                          "bpm_int", "erq_cog", "erq_exp")]

corplot2 <- ggpairs(ggpair_df4, columns = 1:ncol(ggpair_df4),
        lower = list(continuous = wrap(lower_func, method = "lm")),
        title = "Figure 4: Scatter plots and correlations for smoke during pregnancy and s
        axisLabels = "show", columnLabels = colnames(ggpair_df4))
corplot2 <- corplot2 + theme(plot.title = element_text(size = 10))
ggsave("corplot2.jpg", corplot2)

# use ggally for correlations and scatterplot
ggpair_df1 <- data[, c("mom_smoke_pregnant", "cotimean_34wk",
                       "cotimean_pp6mo",
                       "bpm_att", "bpm_ext",
                             "bpm_int")]

ggpairs(ggpair_df1, columns = 1:ncol(ggpair_df1),
        lower = list(continuous = wrap(lower_func, method = "lm")),
        title = "",
        axisLabels = "show", columnLabels = colnames(ggpair_df1))

# use ggally for correlations and scatterplot
ggpair_df2 <- data[, c("momcig", "bpm_att", "bpm_ext",
                             "bpm_int", "erq_cog", "erq_exp")]

ggpairs(ggpair_df2, columns = 1:ncol(ggpair_df2),
        title = "",
        lower = list(continuous = wrap(lower_func, method = "lm")),
        axisLabels = "show", columnLabels = colnames(ggpair_df2))

# simple linear regressions
lm1 <- lm(bpm_att ~ smoke_exposure_duration + mom_smoke_pregnant + mom_smoke_post +
     mom_numcig + momcig, data = data)

summary(lm1)

lm2 <- lm(erq_exp ~ smoke_exposure_duration + mom_smoke_pregnant + mom_smoke_post +
     mom_numcig + momcig, data = data)

summary(lm2)
```