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

Wanyi Chen

Introduction

The main objective of this project is to examine the effects of Smoke During Pregnancy (SDP) / Environmental Tobacco Smoke (ETS) on adolescent self-regulation, substance use, and externalizing through an Exploratory Data Analysis. The data has 98 adolescents and mothers that are randomly selected from a larger data set that is originally collected for a study on smoke avoidance intervention to reduce low-income women's (N=738) smoking, and ETS exposure during pregnancy and children's exposure to ETS in the immediate postpartum period. The data includes: some background information for both the adolescents and mothers (age, race, sex, etc), mother's smoking status, smoke exposures from mother to child, brief problem monitor, emotion regulation responses, parental knowledge responses, and SWAN ratings.

Descriptive Statistics and Missing Data

First, we can briefly look at the summary statistics for some of the variables in the data. Table 1 shows some summary statistics from the parent data (data of the 49 mothers only). We can see that the mean age of the mothers is 38, and most of them white (53%). 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 `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_34wk` is 50 and the mean for `cotimean_pp6mo` is 100, which are very different cotinine levels at and after pregnancy. 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.

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 and after pregnancy are between 24% to 40%, the child's smoking, substance use, and alcohol use percentages are relatively low.

Parent Variable	Parent Missingness (%)	Child Variable	Child Missingness (%)
childasd	57.14	cig_ever	24.49
nidaalc	20.41	num_cigs_30	97.96
nidatob	20.41	e_cig_ever	24.49
nidapres	22.45	num_e_cigs_30	95.92
nidaill	20.41	mj_ever	24.49
momcig	20.41	num_mj_30	93.88
mom_numcig	26.53	alc_ever	26.53
mom_smoke_16wk	2.04	num_alc_30	91.84
mom_smoke_22wk	14.29	bpm_att	24.49
mom_smoke_32wk	18.37	bpm_ext	24.49
mom_smoke_pp1	79.59	bpm_int	28.57
mom_smoke_pp2	40.82	erq_cog	26.53
mom_smoke_pp12wk	14.29	erq_exp	26.53
mom_smoke_pp6mo	18.37	pmq_parental_knowledge	28.57
cotimean_34wk	22.45	pmq_child_disclosure	26.53
cotimean_pp6mo_baby	22.45	pmq_parental_solicitation	28.57
cotimean_pp6mo	22.45	pmq_parental_control	32.65
bpm_att_p	26.53		
bpm_ext_p	24.49		
bpm_int_p	20.41		
smoke_exposure_6mo	20.41		
smoke_exposure_12mo	20.41		
smoke_exposure_2yr	20.41		
smoke_exposure_3yr	22.45		

Parent Variable	Parent Missingness (%)	Child Variable	Child Missingness (%)
smoke_exposure_4yr	22.45		
smoke_exposure_5yr	20.41		
ppmq_parental_knowledge	24.49		
ppmq_child_disclosure	24.49		
ppmq_parental_solicitation	30.61		
ppmq_parental_control	24.49		
bpm_att_a	22.45		
bpm_ext_a	22.45		
bpm_int_a	20.41		
erq_cog_a	20.41		
erq_exp_a	20.41		

Table 3: Calculated missingness percentages for some variables in both the parent and child data.

Then, we can explore the missing patterns in the data. From Table 3, we have the missing percentages for some selected variables in the data. Since we are interested in the effects of SDP/ETS, it is important to examine whether we have sufficient data to support any hypotheses. For the list of parent variables, we see that the variable with highest missing percentage (79.59%) is `mom_smoke_pp1` while other similar variables are mostly ranging from 14% to 18%. 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.

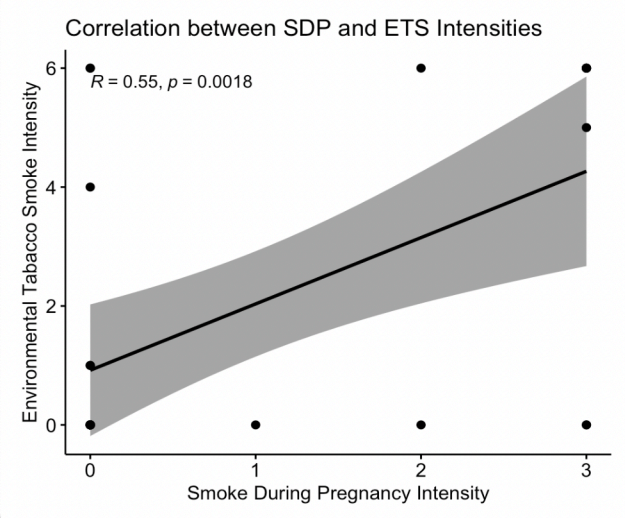

Figure 1: Correlation plot for interrelatedness between prenatal and postnatal smoke exposure.

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.

Correlations

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 a new variable called smoke_exposures_duration to investigate whether higher smoke exposures duration causes poorer ability to handle self-regulation problems for adolescents. smoke_exposures_duration ranges from 0 to 5 years, and is calculated by summing the smoke exposures variables. Then, using this newly created variables, we can plot it against the self-regulation variables and calculate correlations between them. Figure 3 shows the scatter plots of smoke exposure duration and self-regulation variables along with their correlation values on the upper side. In general, all the self-regulation variables are positively correlated with smoke exposures duration. The highest correlation for smoke exposures duration is 0. 537 with erq_exp. This suggests a higher smoke exposures duration is positively correlated with a higher expressive suppression and keeping feeling internally. The other strong and the highest correlation in Figure 3 is 0.658 which is between bpm_ext and bpm_att. 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 smoke exposures duration, they are pretty weak correlarions which could be due to the high number of 0 smoke exposures duration.

In addition to smoke exposures duration, we can also examine the smoking status of mother during pregnancy. We created a new variable called mom_smoke_pregnant by summing the mom smoke status responses during pregnancy period. The variable ranges from 0 to 3, with 0 meaning no smoking at all during pregnancy and 3 indicating 32 weeks of smoking during pregnancy. Using this variable, we can see how mothers' smoking status during pregnancy correlates with the self-regulation variables. From Figure 4, we see that all correlations are also positive. But there are not strong correlations between smoke during pregnancy and self-regulation variables, and only bpm_att (0.323) and erq_exp (0.216) are somewhat correlated with it. From the distribution of mom_smoke_pregnant, we can see that it is very right skewed with most of the mothers not smoking during pregnancy. This makes the interpretation of the correlations difficult since we do not have a lot of smoking 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 3 and Figure 4, correlation does not directly means causation so further analyses are needed, and regressions might be more helpful in seeing the actual effects of SDP/ETS.

Regressions

Based on the correlations we see from Figure 3 and Figure 4, we can also construct simple linear regression model to see whether there is significant effects of SDP/ETS. The first linear regression model summary is shown below. The independent variables are `smoke_exposures_duration`, `mom_smoke_pregnant`, `momcig` (number of cigarettes in the past 30 days for mothers), and `mom_numcig` (number of cigarettes per day for mothers). The dependent variable is `bpm_att`.

From this first simple linear regression model, we can see that there are actually no significant main effects from the independent variables. This is expected since we also do not see very strong positive correlations between them.

The second linear regression model summary is shown below. The independent variables are again`smoke_exposures_duration`, `mom_smoke_pregnant`, `momcig` (number of cigarettes in the past 30 days for mothers), and `mom_numcig` (number of cigarettes per day for mothers). The dependent variable this time is `erq_exp`.

This time, we can see that there is 1 significant main effect from smoke exposures duration. This significant main effect suggests that for an additional unit increase in smoke exposure duration `erq_exp` is expected to increase by 0.3840 unit. This is also expected since we do observe the relative stronger correlation of `erq_exp` and smoke exposures duration in Figure 3. 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 for independent variable selection.

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.

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)
```

```r
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"),]
```

```r
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)
```