Exploratory Data Analysis about Parenatal Tobacco Exposure

Liangkang Wang

Introduction

This analysis is rooted in research conducted by Dr. Lauren Micalizzi, which investigates the impact of smoking during pregnancy (SDP) and environmental tobacco smoke (ETS) exposure on adolescent self-regulation, substance use, and externalizing behaviors. The dataset comprises baseline data of a subset of adolescents and their mothers, with subsequent follow-up assessments planned at 6 and 12 months post-baseline. The aim of this exploratory data analysis is to examine effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing behaviors.

Data Collection

The study involved approximately 800 pregnant individuals as participants. The criteria for participation included being smoke-exposed (current smokers, individuals who quit smoking on their own, or those exposed to the smoke of others), pregnant with only one baby, and having access to a telephone and video player. These participants were randomized into either experimental or control conditions.

- **Intervention Group**: Participants in this group received newsletters containing content aimed at smoking cessation and avoidance, with five newsletters distributed during pregnancy and three after. Additionally, they received videos (three during and two after pregnancy) individually tailored based on behavioral theory-informed survey questions.
- **Comparison Group**: Participants in this group received newsletters and videos focused on healthy pregnancy topics, not specifically related to smoking cessation.

This comprehensive approach to data collection and intervention design forms the basis of the subsequent data analysis.

Variables description

Before dealing with missing value, I want to make sure which variables are important in our following analysis in aim of examining effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing. This part will definitely affect our decision in how to deal with missing values.

(1) Exposure Variables:

We only care about the exposure to smoking during pregnancy (SDP), environmental tobacco exposure (ETS) to mom during pregnancy, and envrionmental tobacco exposre (ETS) to children in the immediate postpartum period. We seperate and define our focused variables below:

SDP-variables:

- `mom_smoke_16wk`, `mom_smoke_22wk`, `mom_smoke_32wk`: Indicate mom smoking status at various pregant periods.
- `cotimean_34wk` : Indicate the Urine continine (nicotine metabolite) at 34 wks gestation.

ETS-variables:

- `mom_smoke_pp1`, `mom_smoke_pp2`, `mom_smoke_pp12wk`, `mom_smoke_pp6mo`: Indicate mom smoking status at various postpartum periods.
- `cotimean_pp6mo`, `cotimean_pp6mo_baby`: Indicate the Urine cotinine (nicotine metabolite) at 6 months postpartum from mom and baby.

(2) Outcome Variables:

Adolescent Self-Regulation: Variables indicating the adolescent's self-regulation ability, which might involve emotional, cognitive, and behavioral control.

- `bpm_att`, `erq_cog`, `erq_exp`, `bpm_att_p`

Substance Use: Variables indicating the adolescent's usage of substances like tobacco, alcohol, or other drugs.

- `cig_ever`, `num_cigs_30`, `e_cig_ever`, `num_e_cigs_30`, `mj_ever`, `num_mj_30`, `alc_ever`, `num_alc_30`

Externalizing Behaviors: Variables indicating behaviors that are outwardly directed and may include hyperactivity, aggression, etc.

- `swan_hyperactive`, `bpm_ext`, `bpm_ext_p`

(3) Demographic Covariates:

children Variables:

- tage, tsex, language, tethnic, taiain, tasian, tnhpi, tblack, twhite, trace_other

Parental/Maternal Variables:

- page, psex, plang, pethnic, paiain, pasian, pnhpi, pblack ,pwhite, prace_other, employ, pedu, income

Missing Data Pattern

In the second part of this report, our objective is to assess the missing values within the dataset. Upon reviewing the dataset, we identified two types of missing values: - NA for numeric columns - "" for character columns.

We initially converted the "" values into NA. Subsequently, we utilized a heatmap to visualize the missing values in the dataset. Within the heatmap, black signifies a missing value. We also transfered the income category from character to numeric value.

The columns with the most significant number of missing values, exceeding 40, are as follows: - num_cigs_30: 48 missing values - num_mj_30: 46 missing values - num_e_cigs_30: 47 missing values - num_alc_30: 45 missing values

These variables all pertain to the substance use category. Conversely, we have another set of variables - e_cig_ever, mg_ever, alc_ever, cigs_ever - that have only 12 or 13 missing values. We are inclined to use these latter variables as outcome variables for subsequent analyses. The rationale being, the num_30 variables offer more detailed and quantifiable data on substance use, while the latter set merely determines prior usage.

Upon examination of our heatmap Figure 1, we observed that there are 10 participants with a substantial number of missing values across multiple columns, including demographic details related to children and their mothers. These participants exhibit a consistent pattern of missing values, as evident from the heatmap. However, crucial variables such as **SDP**, **ETS**, **race**, and swan_hyperactive (indicative of **Externalizing Behaviors**) remain intact for these participants. Considering the fact that we only have 49 observations in our dataset, so we want to include the participants in our study. The critical technique to analyze such a small size data is to subdivide these data into subgroups, which can be used in different univariate analysis targeting our research aim.

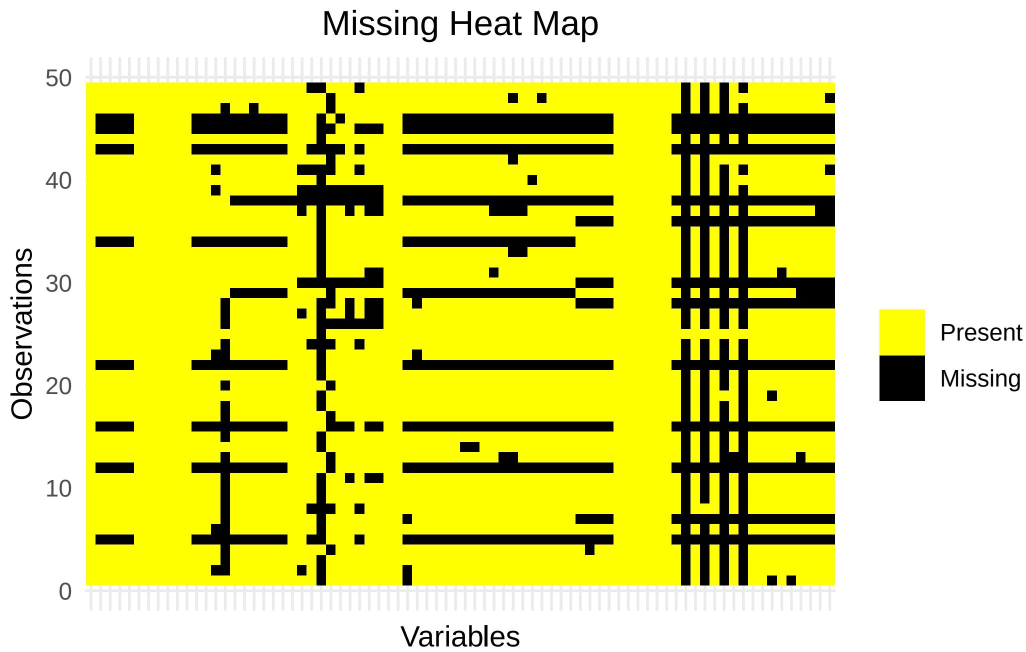


Figure 1: Missing Value Heatmap

Demographic Covariates Distribution

In this part of report, we want to see the population distribution in the research. We can see the distribution of participants' race, age, employment, education, and income.

In the **Age** plot, it's evident that parents' ages, ranging from 32 to 45, exhibit a broader spectrum compared to the children's ages, which fall between 12 and 16. The data originates from a subset of adolescents and their mothers, all of whom were initially part of a previous study focusing on smoke avoidance interventions. While the research exclusively incorporates adolescents, the age at which their mothers gave birth inherently varies, spanning from 18 to 32. It's plausible to consider that a mother's age at childbirth could influence the child's growth and development. Nonetheless, given that our study encompasses only 49 observations, we must operate under the assumption that age does not exert a significant impact, particularly when compared to the notably more severe effects of *SDP* and *ETS*.

In the **Employment** plot, we observe that a preponderance of parents maintain full-time employment, ostensibly presenting a facade of enhanced stability and prosperity compared to their part-time employed counterparts. A plausible inference might be that wealthier families, typically associated with full-time employment, are endowed with more abundant resources, thereby better positioning them to nurture and educate their children. This enhanced capacity for child-rearing could, theoretically, enable children to develop healthier life habits and possibly attain higher scores in self-regulation questionnaires.

However, it's crucial to acknowledge that employment status, and by extension socioeconomic status, could introduce confounding variables into our analysis. The discernible associations between parental employment and children's self-regulation might be influenced, or even obscured, by these underlying socioeconomic factors. For instance, access to superior educational resources, nutritious food, and a stable living environment, often more readily available to wealthier families, can significantly sculpt a child's development and self-regulating capabilities. Consequently, the intricate role of employment status as a potential confounder warrants meticulous consideration and, where possible, statistical adjustment in future analyses to accurately decipher the authentic impact of Smoking During Pregnancy (SDP) and Environmental Tobacco Smoke (ETS) on adolescent outcomes

In the **Race** plot, a notable disparity is evident between the race distributions of parents and children. While certain racial categories, such as American Indian/Alaska Native, White, Hispanic/Latino, and Other, exhibit minor fluctuations between parents and children, substantial variations are present in others.

For instance, while there are 8 Pacific Islander parents recorded, no children are categorized under Pacific Islander. Conversely, despite the absence of Black parents in the data, there are 15 Black children. Furthermore, a significant decline is observed in the count of White individuals from parents to children.

This discrepancy challenges conventional expectations, as children typically share the same racial designation as their parents. A plausible explanation for this phenomenon could stem from situations where parents identify with multiple racial categories, allowing children the flexibility to select their own racial identification. Additionally, the presence of missing values or opting not to provide a race could contribute to this imbalanced data representation.

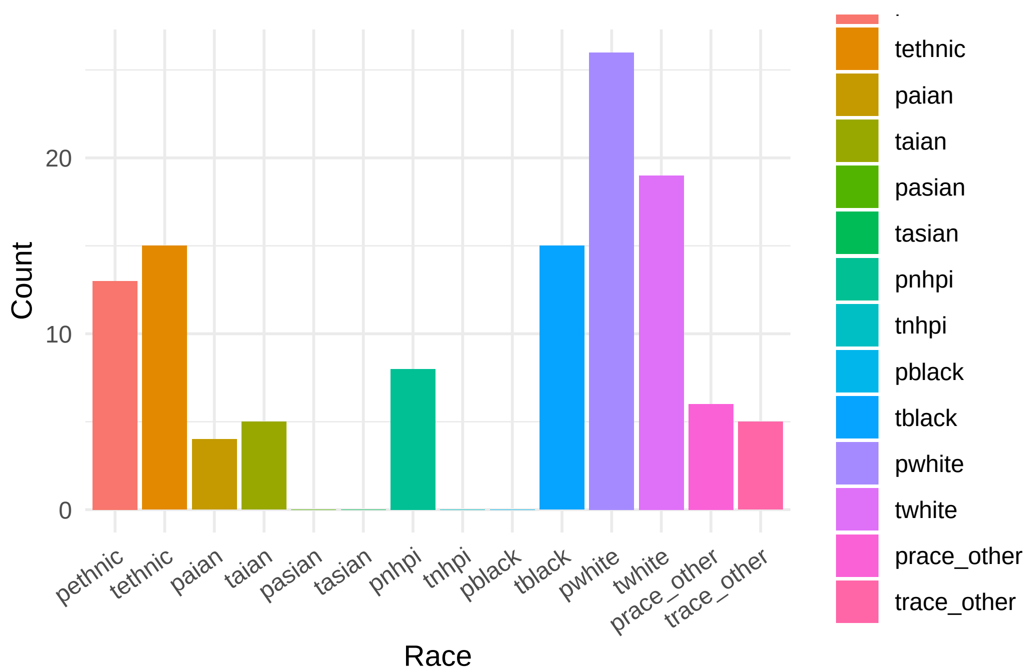
In the **education** plot, it is evident that the predominant educational attainment among parents in our study is "some-college education." This is closely followed by those possessing a 4-year degree, which represents the second-largest category. In contrast, individuals with a postgraduate degree and those with only high school education constitute smaller segments. The distribution of educational levels, with the exception of the 4-year degree category, broadly resembles a normal distribution within our dataset.

We can see the parents **income** appears to follow a uniform distribution if we ignore the 80k-100k group, which has only 1 person. The majority of people in our dataset have an annual income of 40k-60k. However, it's important to note that there are 13 people who do not have an income record. This will undoubtedly affect our parents' income distribution.

Main analysis for effects of SDP/ETS on adolescent self-regulation, substance use, and externalizing

This is the main part of this report. In this part, we will conduct six analyses to detect the effect of SDP/ETS on adolescent self-regulation, substance use, and externalizing. Due to a


Density
0.2
0.1
0.0
20
30
40
Age
Variables
page
tage
Count
20
15
10
5
0
0
1
2
NA
Employment Status
Employment Status
No Job
Part−time
Full−time
Frequency
10
5
0
Income Category
Count
15
10
5
0
Education Status
Education Status
some high school
high school
GED
some college
2 year degree
4 year degree
postgraduate degree


Count
20
10
0
pethnic
tethnic
paian
taian
pasian
tasian
pnhpi
tnhpi
pblack
tblack
pwhite
twhite
prace_other
trace_other
Race
tethnic
paian
taian
pasian
tasian
pnhpi
tnhpi
pblack
tblack
pwhite
twhite
prace_other
trace_other

substantial amount of missing values in the dataset, different subsets of the original data will be utilized for these analyses.

SDP variables preprocessing

As discussed in the variables section, we have mom_smoke_16wk, mom_smoke_22wk, mom_smoke_32wk, and cotimean_34wk to denote **SDP**.

mom_smoke_16wk, mom_smoke_22wk, and mom_smoke_32wk indicate whether moms had smoked during different pregnancy periods, so we aim to collapse them into one variable: Mom_Smoke_p. We will combine these four columns and cotimean_34wk into one table to explore their relationship in terms of missing values. If any smoking variables denoting 1, then we will make the Mom_Smoke_p equals to 1. A strong relationship between Mom_Smoke_p and mom_smoke_32wk was found. Every mom who reported smoking at 32 weeks has a urine cotinine level over 40. We define an imputation method: if cotimean_34wk has a value < 2, and we have a missing value in Mom_Smoke_p, then we will substitute the missing value with no-smoke. Consequently, we have 11 missing values in our SDP variable.

ETS variables preprocessing

Several variables pertain to adolescents' ETS exposure during early postpartum periods, including mom_smoke_pp1, mom_smoke_pp2, mom_smoke_pp12wk, and mom_smoke_pp6mo, which signify the mother's smoking status at distinct postpartum intervals. The implication is straightforward: if a mother smokes, her baby is likely exposed to ETS. The handling of ETS variables mirrors the approach taken with SDP variables.

The variables cotimean_pp6mo and cotimean_pp6mo_baby represent the urine cotinine (a nicotine metabolite) levels at 6 months postpartum for both mother and baby, respectively, serving as quantifiable indicators of the extent to which both were affected by smoking.

Upon analysis, it was observed that cotimean_pp6mo and cotimean_pp6mo_baby share identical patterns of missing values across observations. Furthermore, there is a considerable number of missing values in the postpartum smoking records for mothers. Given that cotinine levels provide a more quantifiable measure, we opted to utilize the cotimean_pp6mo and cotimean_pp6mo_baby variables in lieu of the maternal smoking records.

Adolescent Self-Regulation variables preprocessing

Variables such as bpm_att,bpm_att_p, erq_cog, and erq_exp denote **Adolescent Self-Regulation**.

erq_exp and erq_cog represent the average response on the Emotion Regulation Questionnaire related to Expressive Suppression or Cognitive Reappraisal, respectively. bpm_att and bpm_att_p are the sum of responses on the Brief Problem Monitor on items related to attention problems, both from child self-reports and parent reports for their children. Some correlation between them was also discovered (0.65 of bpm_att and bpm_att_p). Therefore, we plan to use the mean of bpm_att and bpm_att_p to represent the attention problem. If there is one missing value, then we will use the data from the other column. The same approach will be applied to erq_exp and erq_cog for emotion regulation. Finally we get Bpm_att and Erq as our final variables.

Substance Use variables reprocessing

The variables cig_ever, num_cigs_30, e_cig_ever, num_e_cigs_30, mj_ever, num_mj_30, alc_ever, and num_alc_30 are utilized to depict substance use among adolescents. Nevertheless, num_cigs_30, num_e_cigs_30, num_mj_30, and num_alc_30 exhibit a substantial amount of missing values in our dataset. The remaining four variables mirror the substance type represented by these four variables, namely cigarettes, electronic cigarettes, marijuana, and alcohol, respectively. Consequently, we will exclude the four 'num' variables and rely on the alternate four variables for our analysis focused on substance use. These selected variables exhibit a consistent pattern of missing data, facilitating a smoother data analysis process.

Externalizing Behaviors variables reprocessing

The variables swan_hyperactive, bpm_ext, and bpm_ext_p serve to signify behaviors that are externally directed, potentially encompassing hyperactivity and aggression, among others. The swan_hyperactive variable represents the sum of responses on SWAN Rating Scale Questions 1-9, where a score of 6 or higher suggests a likelihood of ADHD - Hyperactive/Impulsive type in the child. Both bpm_ext and bpm_ext_p represent the sum of responses on the Brief Problem Monitor relating to children's externalizing problems. Remarkably, there are no missing values in swan_hyperactive. For bpm_ext and bpm_ext_p, we intend to compute the average of the values and employ imputation for any missing values, maintaining consistency with our prior approach.

Regression model and Analysis

In this part, we made regression models to investigate the associations between SDP/ETS effect and adolescents' behavior, self-Regulation, substance use, and externalizing behaviors. The code and model fitting can be found in appendix.

1. SDP Effect on Adolescent Behavior

We use Mom_Smoke_p, cotimean_34wk to denote the SDP effect, as discussed in the last part.

a. Self-Regulation

Using the `mgcv` package, we investigated the association between SDP indicators (`Mom_Smoke_p` and `cotimean_34wk`) and adolescent self-regulation markers (`Bpm_att` and `Erq`). Key findings include:

- `Mom_Smoke_p` was significantly associated with `Erq` ($p = 0.0182$). Adolescents with mothers who smoked during pregnancy had, on average, an `Erq` score 0.8341 units higher.
- `cotimean_34wk` did not show a significant relationship with `Erq` ($p = 0.146$).

b. Substance Use

We examined the impact of SDP on adolescent substance use:

- **Cigarette Use (`cig_ever`)**: Both maternal smoking and cotinine levels were significantly related ($p = 0.0336$ and $p < 2e\text{-}16$ respectively).
- **E-cigarette Use (`e_cig_ever`)**: Both predictors were significantly associated, explaining 76.3% of the variance.
- **Marijuana Use (`mj_ever`)**: Only cotinine levels were significant ($p = 0.0039$), explaining 60% of the variance.
- **Alcohol Use (`alc_ever`)**: Neither predictor was significant, accounting for 23.8% variance.

c. Externalizing Behaviors

Exploring associations with adolescent hyperactivity (`swan_hyperactive`) and externalizing behaviors (`Bpm_ext`):

- **Hyperactivity**: Neither SDP predictor was significant.
- **Externalizing Behaviors**: Both predictors were not significantly related.

2. ETS Effect on Adolescent Behavior

We chose to utilize the `cotimean_pp6mo` and `cotimean_pp6mo_baby` variables to denote the ETS effect on children.

a. Self-Regulation

We assessed the associations between ETS exposure (`cotimean_pp6mo` and `cotimean_pp6mo_baby`) and self-regulation:

- **Self-Regulation (`Bpm_att`)**: Only maternal cotinine levels at 6 months postpartum were significant ($p = 0.0334$).
- **Emotional Regulation (`Erq`)**: Neither ETS predictor was significant.

b. Substance Use

Investigating the relationship between ETS exposure and adolescent substance use:

- **Cigarette Use (`cig_ever`)**: Only maternal cotinine levels were significant ($p = 0.00518$).
- **E-cigarette Use (`e_cig_ever`)**: Neither predictor was significant.
- **Marijuana Use (`mj_ever`)**: Both predictors were not significant.
- **Alcohol Use (`alc_ever`)**: Only maternal cotinine levels were significant ($p = 0.000357$).

c. Externalizing Behaviors

Examining ETS effects on adolescent externalizing behaviors:

- **Hyperactivity (`swan_hyperactive`)**: Only maternal cotinine levels were significant ($p = 0.0489$).
- **Externalizing Behaviors (`Bpm_ext`)**: Neither predictor was significant.

Conclusion

Our analysis provides insight into the potential risks associated with maternal smoking during pregnancy and environmental tobacco exposure postpartum on adolescent behaviors. The findings indicate a complex interplay between these factors and various adolescent behaviors and outcomes.

Emotional and Behavioral Regulation: SDP, particularly maternal smoking, seems to elevate the risk for issues in adolescent emotional regulation. This is indicative of the lasting impact that prenatal exposure to smoking can have on the neurodevelopmental processes related to emotional and behavioral control. Substance Use: The effects of SDP and ETS exposure on substance use among adolescents are varied but significant. Our data suggests that SDP increases the likelihood of adolescents engaging in the use of cigarettes, e-cigarettes, and marijuana. Similarly, ETS exposure postpartum correlates with increased usage of cigarettes, e-cigarettes, and alcohol among adolescents. These findings underscore the broader implications of tobacco exposure during critical developmental periods on the propensity for substance use in later life.

Postpartum Environmental Exposure: Interestingly, ETS exposure postpartum, particularly reflected in maternal cotinine levels, shows more pronounced effects on certain behaviors like cigarette use and hyperactivity. This highlights the potential influence of postnatal environmental factors on adolescent behavior, beyond the impact of prenatal exposure.

The results of our analysis, while indicative, are not conclusive. They underscore the complexity of the relationship between maternal smoking, environmental tobacco exposure, and adolescent developmental outcomes. It's clear that both prenatal and postnatal exposures contribute to a range of behavioral and emotional challenges in adolescents, but the exact nature and extent of these effects vary.

Given the intricate nature of these relationships, further studies are crucial. Longitudinal research that tracks individuals from prenatal stages through adolescence and into adulthood could provide more comprehensive insights into these dynamics. Additionally, exploring potential interventions and preventive measures is essential. Understanding the mechanisms through which maternal smoking and environmental tobacco exposure impact adolescent development can inform targeted strategies to mitigate these risks. Such interventions could play a crucial role in breaking the cycle of intergenerational transmission of health and behavioral problems associated with tobacco exposure.

In summary, our analysis contributes to the growing body of evidence on the negative impacts of maternal smoking and environmental tobacco exposure. It underscores the need for heightened awareness and interventions aimed at reducing tobacco exposure during these critical developmental periods.

Model Fit Appendix:

SDP effect on Adolescent Self-Regulation

```
Family: gaussian
Link function: identity

Formula:
Bpm_att ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0209     0.5481   5.512  1.1e-05 ***
Mom_Smoke_p  -1.1537     1.2741  -0.905    0.374
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                   edf Ref.df     F p-value
s(cotimean_34wk) 6.713   7.63 1.738    0.13

R-sq.(adj) =  0.272   Deviance explained = 44.8%
GCV = 5.3814  Scale est. = 3.9605    n = 33
```

```
Family: gaussian
Link function: identity

Formula:
Erq ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6879     0.1668  16.115 1.07e-15 ***
Mom_Smoke_p   0.8341     0.3325   2.509   0.0182 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df     F p-value
s(cotimean_34wk)   1      1 2.235   0.146
```

```
R-sq.(adj) =  0.127   Deviance explained = 18.5%
GCV = 0.47725  Scale est. = 0.43107   n = 31
```

SDP effect on Adolescent Substance Use

```
Model: Predicting cig_ever

Family: gaussian
Link function: identity

Formula:
cig_ever ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 3.333e-02  2.620e-08 1.272e+06   <2e-16 ***
Mom_Smoke_p 1.387e-07  6.056e-08 2.290e+00   0.0336 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df         F p-value
s(cotimean_34wk)   9      9 1.195e+13  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =      1   Deviance explained =  100%
GCV = 1.3212e-14  Scale est. = 8.3676e-15  n = 30

Model: Predicting e_cig_ever

Family: gaussian
Link function: identity

Formula:
e_cig_ever ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.007743   0.040585  -0.191   0.8505
```

```
Mom_Smoke_p  0.223228   0.092865   2.404   0.0253 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                   edf Ref.df     F  p-value
s(cotimean_34wk) 6.563  7.513 7.058 0.000188 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.679   Deviance explained = 76.3%
GCV = 0.028925  Scale est. = 0.020669  n = 30

Model: Predicting mj_ever

Family: gaussian
Link function: identity

Formula:
mj_ever ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.11165    0.06165   1.811   0.0835 .
Mom_Smoke_p -0.03494    0.14078  -0.248   0.8062
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                   edf Ref.df     F p-value
s(cotimean_34wk) 5.499  6.428 4.286  0.0039 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.485   Deviance explained =   60%
GCV = 0.063934  Scale est. = 0.047953  n = 30

Model: Predicting alc_ever

Family: gaussian
Link function: identity

Formula:
```

```
alc_ever ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14100    0.09233   1.527    0.139
Mom_Smoke_p  0.07700    0.20103   0.383    0.705

Approximate significance of smooth terms:
                   edf Ref.df     F p-value
s(cotimean_34wk) 1.781  2.168 2.458   0.113

R-sq.(adj) =  0.158   Deviance explained = 23.8%
GCV = 0.13849  Scale est. = 0.12104   n = 30
```

SDP effect on Adolescent Externalizing Behaviors

```
Model: Predicting swan_hyperactive

Family: gaussian
Link function: identity

Formula:
swan_hyperactive ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.020      1.456   4.136 0.000263 ***
Mom_Smoke_p    4.396      3.026   1.453 0.156712
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                 edf Ref.df     F p-value
s(cotimean_34wk)   1      1 1.054   0.313

R-sq.(adj) =  0.192   Deviance explained = 24.2%
GCV = 39.971  Scale est. = 36.337    n = 33

Model: Predicting Bpm_ext

Family: gaussian
```

```
Link function: identity

Formula:
Bpm_ext ~ Mom_Smoke_p + s(cotimean_34wk)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.8525     0.5265   5.418 7.78e-06 ***
Mom_Smoke_p  -1.0120     1.1590  -0.873     0.39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                   edf Ref.df     F p-value
s(cotimean_34wk) 1.809    2.2 1.193   0.331

R-sq.(adj) =  0.0228   Deviance explained = 10.9%
GCV = 4.7722  Scale est. = 4.2214    n = 33
```

ETS effect on Adolescent Self-Regulation

```
Model: Predicting Bpm_att

Family: gaussian
Link function: identity

Formula:
Bpm_att ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5323     0.3063   8.268 2.05e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      2.916  3.496 3.286  0.0334 *
s(cotimean_pp6mo_baby) 3.572  4.303 1.761  0.1671
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
R-sq.(adj) =  0.361   Deviance explained = 49.9%
GCV = 3.8344  Scale est. = 2.9082    n = 31

Model: Predicting Erq

Family: gaussian
Link function: identity

Formula:
Erq ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9530     0.1272   23.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      1.000  1.000 0.000   0.983
s(cotimean_pp6mo_baby) 1.117  1.224 0.125   0.886

R-sq.(adj) =  -0.0637   Deviance explained = 1.13%
GCV = 0.55788  Scale est. = 0.50179   n = 31
```

ETS effect on Adolescent Substance Use

```
Model: Predicting cig_ever

Family: gaussian
Link function: identity

Formula:
cig_ever ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03333    0.02769   1.204    0.239

Approximate significance of smooth terms:
```

```
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      1.992  2.433 5.784 0.00518 **
s(cotimean_pp6mo_baby) 1.000  1.000 0.019 0.89079
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.31   Deviance explained = 38.1%
GCV = 0.026526  Scale est. = 0.022996  n = 30

Model: Predicting e_cig_ever

Family: gaussian
Link function: identity

Formula:
e_cig_ever ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1000     0.0556   1.799   0.0833 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                       edf Ref.df     F p-value
s(cotimean_pp6mo)        1      1 2.113   0.158
s(cotimean_pp6mo_baby)   1      1 0.017   0.896

R-sq.(adj) =  0.00392   Deviance explained = 7.26%
GCV = 0.10304  Scale est. = 0.092739  n = 30

Model: Predicting mj_ever

Family: gaussian
Link function: identity

Formula:
mj_ever ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10000    0.05463    1.83   0.0784 .
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      1.000   1.00 1.474   0.235
s(cotimean_pp6mo_baby) 1.377   1.66 0.254   0.644

R-sq.(adj) =  0.0382   Deviance explained = 11.7%
GCV = 0.10091  Scale est. = 0.089547  n = 30

Model: Predicting alc_ever

Family: gaussian
Link function: identity

Formula:
alc_ever ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1667     0.0472   3.531  0.00169 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df    F  p-value
s(cotimean_pp6mo)      3.801  4.561 7.40 0.000357 ***
s(cotimean_pp6mo_baby) 1.000  1.000 0.15 0.702173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.535   Deviance explained = 61.2%
GCV = 0.082861  Scale est. = 0.066838  n = 30
```

ETS effect on Adolescent Externalizing Behaviors

```
Model: Predicting swan_hyperactive

Family: gaussian
Link function: identity
```

```
Formula:
swan_hyperactive ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.727      1.023   7.551 2.41e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      1.785  2.183 3.268  0.0489 *
s(cotimean_pp6mo_baby) 1.000  1.000 0.396  0.5340
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.161   Deviance explained = 23.4%
GCV = 39.037  Scale est. = 34.56     n = 33

Model: Predicting Bpm_ext

Family: gaussian
Link function: identity

Formula:
Bpm_ext ~ s(cotimean_pp6mo) + s(cotimean_pp6mo_baby)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4394     0.3633   6.714 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                         edf Ref.df     F p-value
s(cotimean_pp6mo)      1.138  1.262 0.245   0.623
s(cotimean_pp6mo_baby) 1.000  1.000 0.318   0.577

R-sq.(adj) =  -0.0375   Deviance explained = 3.19%
GCV = 4.8145  Scale est. = 4.3567    n = 33
```

Code Appendix:

```
library(knitr)
library(tidyr)
library(formatR)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(warning = FALSE, message = FALSE, echo = FALSE, fig.align = "center"
library(tableone)
library(tidyverse)
library(reshape)
library(mgcv)
source(file = "Utils.R")
data <- read.csv("project1_data/project1.csv")
data$income[data$income==""]=NA
data$mom_numcig[data$mom_numcig==""]=NA
data$mom_smoke_16wk[data$mom_smoke_16wk==""]=NA
data$mom_smoke_22wk[data$mom_smoke_22wk==""]=NA
data$mom_smoke_32wk[data$mom_smoke_32wk==""]=NA
data$mom_smoke_pp1[data$mom_smoke_pp1==""]=NA
data$mom_smoke_pp2[data$mom_smoke_pp2==""]=NA
data$mom_smoke_pp12wk[data$mom_smoke_pp12wk==""]=NA
data$mom_smoke_pp6mo[data$mom_smoke_pp6mo==""]=NA

data$income=as.numeric(data$income)
# data_long <- data %>%
#   mutate(row_id = row_number()) %>%  # Create a row id
#   gather(key = "variable", value = "value", -row_id) %>%  # Convert to long format
#   mutate(is_missing = is.na(value))  # Create a logical column for missing data
#
# ggplot(data_long, aes(x = variable, y = row_id)) +
#   geom_tile(aes(fill = is_missing), color = "white") +
#   scale_fill_manual(values = c("white", "blue"), name = "Missing Value") +
#   theme_minimal() +
#   labs(x = "Variable", y = "Row ID") +
#   theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())  # Hide x-axis te
#
# # Count NAs in each column
# na_count_cols <- apply(data, 2, function(x) sum(is.na(x)))
#
# # Count NAs in each row
```

```r
# na_count_rows <- apply(data, 1, function(x) sum(is.na(x)))

missing_heatmap(data = data,title = "Missing Heat Map")+
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank())

# Plot function define
variable_den <- function(data_melted,xtitle){
  ggplot(data_melted, aes(x=value, fill=variable)) +
  geom_density(alpha=0.5) +
  labs(x=xtitle,
       y="Density",
       fill="Variables") +
  theme_minimal()
}

# Age
ptage <- data[, c("page", "tage")] %>% melt() %>% filter(!is.na(value))
p_age <- variable_den(data_melted = ptage,xtitle = "Age")

# Employment

employ <- data[, c("employ")] %>% melt() %>% filter(!is.na(value))

# Plot
p_employ <- ggplot(data, aes(x=factor(employ))) +
  geom_bar(aes(y = ..count.., fill = factor(employ)), stat="count") +
  scale_fill_manual(values=c("blue", "red", "green"),
                    name="Employment Status",
                    breaks=c("0", "1", "2"),
                    labels=c("No Job", "Part-time", "Full-time")) +
  labs(x="Employment Status",
       y="Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))


#### This is for the bar plot.
# Race

# Select and reshape relevant columns for ggplot
parent_race_cols <- c("pethnic","paian", "pasian", "pnhpi", "pblack", "pwhite","prace_othe
```

```
child_race_cols <- c("tethnic","taian", "tasian", "tnhpi", "tblack", "twhite","trace_other
# Replace with actual column names

# Melt the data and create a column for race
parent_race <- melt(data, id.vars = "parent_id", measure.vars = parent_race_cols) %>% filt
child_race <- melt(data, id.vars = "parent_id", measure.vars = child_race_cols) %>% filter

# Add a column to indicate whether the record belongs to a child or parent
parent_race$demographic <- "Parent"
child_race$demographic <- "Child"

# Combine the data into a single data frame
all_race <- rbind(parent_race, child_race)
#all_race <- rbind(all_race,c(1000,"tnhpi",0))
#all_race <- rbind(all_race,c(1001,"pblack",0))

# Define custom colors and order
custom_order <- c("pethnic","tethnic","paian","taian", "pasian","tasian", "pnhpi","tnhpi",

all_race$variable <- factor(all_race$variable,levels = custom_order)

race_count <- all_race%>%group_by(variable)%>%count()
variable <- c("tnhpi","pblack","pasian","tasian")
n <- rep(0,4)

B <- race_count%>% as.data.frame()
race_count.new <- rbind(B,cbind(variable,n))
race_count.new$n <- as.numeric(race_count.new$n)

# Plot
p_race <- ggplot(race_count.new, aes(x=variable,y=n,fill=variable)) +
  geom_bar(stat="identity",position="dodge") +
  labs(x="Race",
       y="Count",
       fill="Race Group") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 35, hjust = 1))



# Education
```

```r
education <- data[, c("pedu")] %>% melt() %>% filter(!is.na(value))

# Plot
p_edu <- ggplot(data, aes(x=factor(pedu))) +
  geom_bar(aes(y = ..count.., fill = factor(pedu)), stat="count") +
  scale_fill_manual(values=c("blue", "red", "green","yellow","purple","orange","lightgreen
                    name="Education Status",
                    breaks=c("0", "1", "2","3","4","5","6"),
                    labels=c("some high school", "high school", "GED","some college","2 ye
  labs(x="Education Status",
       y="Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 0, hjust = 1))


# Convert income to numeric, handling non-numeric entries and missing values
Income=data[,c("parent_id","income")]
Income$income <- as.numeric(as.character(data$income))

# Categorizing income
Income$income_category <- cut(Income$income,
                              breaks = c(-Inf, 20000, 40000, 60000, 80000, 100000, Inf),
                              labels = c("<20K", "20K-40K", "40K-60K", "60K-80K", "80K-100K"
                              right = FALSE,
                              include.lowest = TRUE)

# Handling NA values if needed
Income$income_category <- factor(Income$income_category, levels = c("<20K", "20K-40K", "40

# Creating the plot
p_income <- ggplot(Income, aes(x = income_category)) +
  geom_bar(fill = "skyblue", color = "black") +
  labs(x = "Income Category",
       y = "Frequency") +
  theme_minimal()


library(cowplot)
# p_age
# p_employ
# p_race
```

```
# p_edu
# p_income

combined_plot <- plot_grid(
  plot_grid(p_age, p_employ, ncol = 2), # First row with 2 plots
  plot_grid(p_income, p_edu, ncol = 2), # Second row with 2 plots
  ncol = 1, # Ensure the rows are stacked vertically
  align = 'v', # Align vertically
  rel_heights = c(1, 1, 1) # Equal heights for each row
)
combined_plot
p_race
# make new variable of Mom_Smoke_xxwk,with numeric data.
data$Mom_Smoke_16wk=NA
data$Mom_Smoke_16wk[data$mom_smoke_16wk=="1=Yes"]=1
data$Mom_Smoke_16wk[data$mom_smoke_16wk=="2=No"]=0

data$Mom_Smoke_22wk=NA
data$Mom_Smoke_22wk[data$mom_smoke_22wk=="1=Yes"]=1
data$Mom_Smoke_22wk[data$mom_smoke_22wk=="2=No"]=0

data$Mom_Smoke_32wk=NA
data$Mom_Smoke_32wk[data$mom_smoke_32wk=="1=Yes"]=1
data$Mom_Smoke_32wk[data$mom_smoke_32wk=="2=No"]=0

data$Mom_Smoke_p=NA
data$Mom_Smoke_p[data$Mom_Smoke_16wk==1 | data$Mom_Smoke_22wk==1|data$Mom_Smoke_32wk==1]=1
data$Mom_Smoke_p[data$Mom_Smoke_16wk==0 & data$Mom_Smoke_22wk==0 & data$Mom_Smoke_32wk==0]

### if some missing value exist, but there are 0, then we will say it 0. Because most moms
data$Mom_Smoke_p[data$cotimean_34wk<2 & is.na(data$Mom_Smoke_p)]=0

data.mom_smoke <- data[,c("Mom_Smoke_16wk","Mom_Smoke_22wk","Mom_Smoke_32wk","Mom_Smoke_p"
# make new variable of Mom_Smoke_xxwkpp,with numeric data.
data$Mom_Smoke_pp1=NA
data$Mom_Smoke_pp1[data$mom_smoke_pp1=="1=Yes"]=1
data$Mom_Smoke_pp1[data$mom_smoke_pp1=="2=No"]=0

data$Mom_Smoke_pp2=NA
data$Mom_Smoke_pp2[data$mom_smoke_pp2=="1=Yes"]=1
data$Mom_Smoke_pp2[data$mom_smoke_pp2=="2=No"]=0
```

```
data$Mom_Smoke_pp12wk=NA
data$Mom_Smoke_pp12wk[data$mom_smoke_pp12wk=="1=Yes"]=1
data$Mom_Smoke_pp12wk[data$mom_smoke_pp12wk=="2=No"]=0

data$Mom_Smoke_pp6mo=NA
data$Mom_Smoke_pp6mo[data$mom_smoke_pp6mo=="1=Yes"]=1
data$Mom_Smoke_pp6mo[data$mom_smoke_pp6mo=="2=No"]=0

data$Mom_Smoke_pp=NA
data$Mom_Smoke_pp[data$Mom_Smoke_pp1==1 | data$Mom_Smoke_pp2==1|data$Mom_Smoke_pp12wk==1 |
data$Mom_Smoke_pp[data$Mom_Smoke_pp1==0 & data$Mom_Smoke_pp2==0 & data$Mom_Smoke_pp12wk==0


data_ms_pp <- data[,c("Mom_Smoke_pp1","Mom_Smoke_pp2","Mom_Smoke_pp12wk","Mom_Smoke_pp6mo"

#cor(data$bpm_att,data$bpm_att_p,use = "pairwise.complete.obs")
#cor(data$erq_cog,data$erq_exp,use = "pairwise.complete.obs")

data$Bpm_att=NA
data$Bpm_att=(data$bpm_att+data$bpm_att_p)/2
data$Bpm_att[is.na(data$Bpm_att)]=data$bpm_att[is.na(data$Bpm_att)]
data$Bpm_att[is.na(data$Bpm_att)]=data$bpm_att_p[is.na(data$Bpm_att)]

data$Erq=NA
data$Erq=(data$erq_cog+data$erq_exp)/2
data$Erq[is.na(data$Erq)]=data$erq_cog[is.na(data$Erq)]
data$Erq[is.na(data$Erq)]=data$erq_exp[is.na(data$Erq)]

data_sereg <- data[,c("Bpm_att","bpm_att","bpm_att_p","Erq","erq_cog","erq_exp")]
data_substance <- data[,c("cig_ever","e_cig_ever","mj_ever","alc_ever")]
#data$swan_hyperactive
#data$bpm_ext
#data$bpm_ext_p

data$Bpm_ext=NA
data$Bpm_ext=(data$bpm_ext+data$bpm_ext_p)/2
data$Bpm_ext[is.na(data$Bpm_ext)]=data$bpm_ext[is.na(data$Bpm_ext)]
data$Bpm_ext[is.na(data$Bpm_ext)]=data$bpm_ext_p[is.na(data$Bpm_ext)]

data_ext <- data[,c("Bpm_ext","bpm_ext","bpm_ext_p")]
data_sdp_reg <- data[,c("Bpm_att","Erq","Mom_Smoke_p","cotimean_34wk")]
```

```r
data_sdp_reg_1_clean <- na.omit(data_sdp_reg[,-2])
# Fit a GAM model
model_gam_sdp_reg_1 <- gam(Bpm_att ~ Mom_Smoke_p + s(cotimean_34wk), data = data_sdp_reg_1
summary(model_gam_sdp_reg_1)

# Plotting the smooth terms in the model
#plot(model_gam_sdp_reg_1, se = TRUE, pages = 1)

# Predict and plot the results
pred <- predict(model_gam_sdp_reg_1, newdata = data_sdp_reg_1_clean, type = "response", se

# Adding predictions to the original data
data_sdp_reg_1_clean$pred <- pred$fit

# Plotting the original and predicted values
# p_sdp_reg_1 <- ggplot(data_sdp_reg_1_clean, aes(x = cotimean_34wk)) +
#   geom_point(aes(y = Bpm_att)) +
#   geom_line(aes(y = pred), color = "blue") +
#   labs(y = "Bpm_att", x = "cotimean_34wk", title = "GAM Model Predictions vs. Original D
#   theme_minimal()


data_sdp_reg <- data[,c("Bpm_att","Erq","Mom_Smoke_p","cotimean_34wk")]
data_sdp_reg_2_clean <- na.omit(data_sdp_reg[,-1])
# Fit a GAM model
model_gam_sdp_reg_2 <- gam(Erq ~ Mom_Smoke_p + s(cotimean_34wk), data = data_sdp_reg_2_cle
summary(model_gam_sdp_reg_2)

# Plotting the smooth terms in the model
#plot(model_gam_sdp_reg_1, se = TRUE, pages = 1)

# Predict and plot the results
pred <- predict(model_gam_sdp_reg_2, newdata = data_sdp_reg_2_clean, type = "response", se

# Adding predictions to the original data
data_sdp_reg_2_clean$pred <- pred$fit

# Plotting the original and predicted values
# p_sdp_reg_2 <- ggplot(data_sdp_reg_2_clean, aes(x = cotimean_34wk)) +
#   geom_point(aes(y = Erq)) +
#   geom_line(aes(y = pred), color = "blue") +
```

```
#   labs(y = "Erq", x = "cotimean_34wk", title = "GAM Model Predictions vs. Original Data"
#   theme_minimal()

data_sdp_sub <- data[,c("Mom_Smoke_p","cotimean_34wk","cig_ever","e_cig_ever","mj_ever","a
data_sdp_sub_clean <- na.omit(data_sdp_sub)

# Specify the dependent variables
indep_vars <- c("Mom_Smoke_p","cotimean_34wk")

# Specify the independent binary variables
dep_vars <- c("cig_ever", "e_cig_ever", "mj_ever", "alc_ever")

# Loop through each independent variable and create a model for each dependent variable
for (dep_var in dep_vars) {
    # Formulate the model formula
    model_formula <- as.formula(paste(dep_var, "~ Mom_Smoke_p + s(cotimean_34wk)"))
    # Fit the model
    model <- gam(model_formula, data = data_sdp_sub_clean)
    # Display the summary
    cat("\nModel: Predicting", dep_var,"\n")
    print(summary(model))
}


data_sdp_ext <- data[,c("Mom_Smoke_p","cotimean_34wk","swan_hyperactive","Bpm_ext")]
data_sdp_ext_clean <- na.omit(data_sdp_ext)

# Specify the independent binary variables
dep_vars <- c("swan_hyperactive","Bpm_ext")

# Loop through each independent variable and create a model for each dependent variable
for (dep_var in dep_vars) {
    # Formulate the model formula
    model_formula <- as.formula(paste(dep_var, "~ Mom_Smoke_p + s(cotimean_34wk)"))
    # Fit the model
    model <- gam(model_formula, data = data_sdp_ext_clean)
    # Display the summary
    cat("\nModel: Predicting", dep_var,"\n")
    print(summary(model))
}
```

```
data_ets_reg <- data[,c("cotimean_pp6mo","cotimean_pp6mo_baby","Bpm_att","Erq")]
data_ets_reg_clean <- na.omit(data_ets_reg)

# Specify the independent binary variables
dep_vars <- c("Bpm_att","Erq")

# Loop through each independent variable and create a model for each dependent variable
for (dep_var in dep_vars) {
    # Formulate the model formula
    model_formula <- as.formula(paste(dep_var, "~ s(cotimean_pp6mo) + s(cotimean_pp6mo_bab
    # Fit the model
    model <- gam(model_formula, data = data_ets_reg_clean)
    # Display the summary
    cat("\nModel: Predicting", dep_var,"\n")
    print(summary(model))
}



data_ets_sub <- data[,c("cotimean_pp6mo","cotimean_pp6mo_baby","cig_ever","e_cig_ever","mj
data_ets_sub_clean <- na.omit(data_ets_sub)

# Specify the independent binary variables
dep_vars <- c("cig_ever","e_cig_ever","mj_ever","alc_ever")

# Loop through each independent variable and create a model for each dependent variable
for (dep_var in dep_vars) {
    # Formulate the model formula
    model_formula <- as.formula(paste(dep_var, "~ s(cotimean_pp6mo) + s(cotimean_pp6mo_bab
    # Fit the model
    model <- gam(model_formula, data = data_ets_sub_clean)
    # Display the summary
    cat("\nModel: Predicting", dep_var,"\n")
    print(summary(model))
}



data_ets_ext <- data[,c("cotimean_pp6mo","cotimean_pp6mo_baby","swan_hyperactive","Bpm_ext
data_ets_ext_clean <- na.omit(data_ets_ext)
```

```r
# Specify the independent binary variables
dep_vars <- c("swan_hyperactive","Bpm_ext")

# Loop through each independent variable and create a model for each dependent variable
for (dep_var in dep_vars) {
    # Formulate the model formula
    model_formula <- as.formula(paste(dep_var, "~ s(cotimean_pp6mo) + s(cotimean_pp6mo_bab
    # Fit the model
    model <- gam(model_formula, data = data_ets_ext_clean)
    # Display the summary
    cat("\nModel: Predicting", dep_var,"\n")
    print(summary(model))
}
```