ECON 570 Big Data Econometrics

Final Project Group 5

The Analysis of Casual Effects of Mother's Smoking on Baby's Birth Weights

Yingni Fan: 2248623030

Yiqi Zang: 6645326366

Yuqi Pang: 8920670191

Yuchen Hu: 8196508139

Yidan Chen: 6505131988

1 Introduction

The harm of smoking is widely known by the population, which can cause lung, heart disease, and many other diseases. Furthermore, when it is associated with pregnancy, tobacco would not only damage the health of pregnant women but also lead to fetal health and viability problems. Medical research shows that maternal smoking during pregnancy can affect placental blood flow and vascular resistance, leading to impaired nutrient and oxygen transport from the mother to the fetus, which may lead to a decrease in the baby's birth weight because of malnutrition. The effects of maternal smoking during pregnancy on a baby's birth weight have been studied by scholars in different fields for many years, like economics, medicine, and sociology.

Carter (2006) pointed out that smoking lowered mean birth weight between 149.2 grams to 204.3 grams, but there is no significant relationship between smoking and preterm birth. And the harm of maternal smoking accumulates throughout the period of gestation (Hebal et al 1988). Kabir Dasgupta, Keshar M. Ghimire, and Gail Pacheco (2019) found out that the negative effects might continue through the preschool years. This paper is written to evaluate the causal effect of a mother's smoking on her baby's birth weights, based on Natality Data Sets between 1990 and 1998 from the National Center for Health Statistics. Since the true causal effect β is unknown, it is difficult to assess whether our estimated β-hat is close to its true value. Thus, four different identification strategies are constructed to evaluate the causal effect. We will compare the causal effects estimated by those four methods, and then give a possible range of the causal effect.

2 Description of Data

Following are graphs of data summary and the distribution of variables:

Table 1: Summary of data

Variable	Obs	Mean	Std. Dev.	Min	Max
momid3	283,858	70965	40971.45	1	141929
idx	283,858	1.5	0.500001	1	2
stateres	283,858	26.60711	14.59492	1	51
dmage	283,858	28.37494	5.435081	13	50
dmeduc	283,858	13.88196	2.270343	0	17
mplbir	283,858	26.14396	14.33209	1	51
nlbnl	283,858	1.18826	1.203141	0	15
gestat	283,858	39.25423	2.165894	17	47
dbirwt	283,858	3454.168	539.1527	227	8020
cigar	283,858	1.630111	5.01608	0	99
smoke	283,858	0.130114	0.33643	0	1
male	283,858	0.513295	0.499824	0	1
year	283,858	3.854864	2.259622	0	8
married	283,858	0.868906	0.337503	0	1
hsgrad	283,858	0.294986	0.456037	0	1
somecoll	283,858	0.234336	0.423584	0	1
collgrad	283,858	0.377773	0.484831	0	1
agesq	283,858	834.6771	311.1107	169	2500
black	283,858	0.074263	0.262198	0	1
adeqcode2	283,858	0.167538	0.373456	0	1
adeqcode3	283,858	0.038836	0.193205	0	1
novisit	283,858	0.007218	0.084654	0	1
pretri2	283,858	0.111922	0.315271	0	1
pretri3	283,858	0.01957	0.138516	0	1

Table 2: Distribution of data

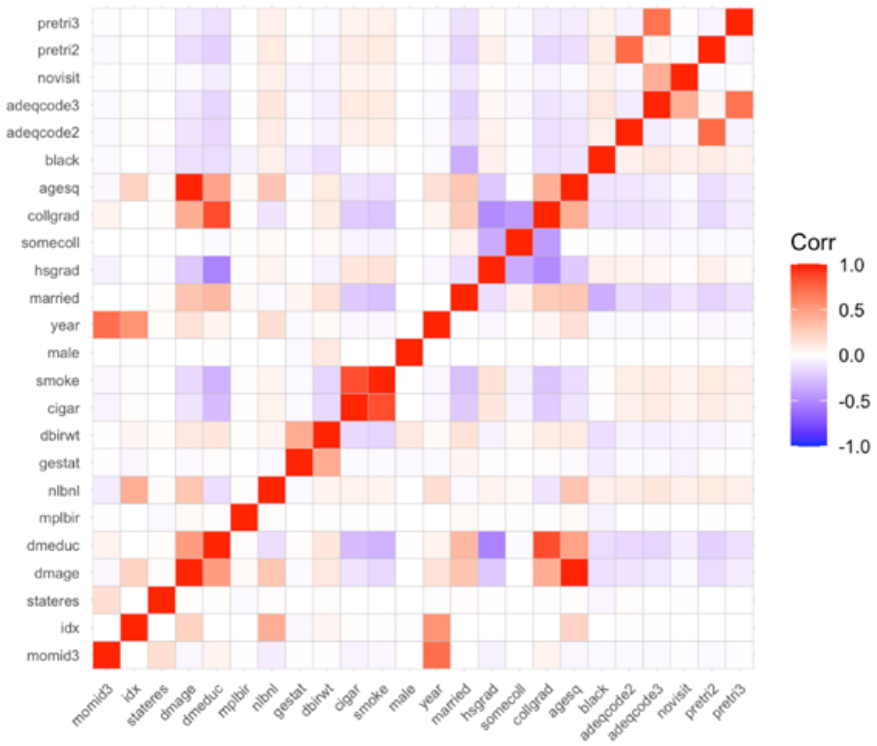
Variable		Number	Percentage
Baby's birth weights (grams)	＜2500	11014	3.9%
	≥2500	272844	96.1%
Age (years)	≤19	14768	5.2%
	20-34	230316	81.1%
	≥35	38774	13.7%
Education (years)	≤8	3638	1.3%
	9-11	22734	8%
	≥12	257486	90.7%
Length of gestation (weeks)	≤37	20363	7.2%
	38-42	251690	88.7%
	≥43	11805	4.1%

Education level	High-school graduate	83734	29.5%
	Some-college	66518	23.4%
	College-graduate	107234	37.8%
Smoke	No	246924	87%
	Yes, 1-9 cigarettes per day	10093	3.6%
	Yes, 10-20 cigarettes per day	23369	8.2%
	Yes, 21-98 cigarettes per day	1890	0.7%
	Yes, but the number is unsure	1643	0.5%
Married	Yes	246646	86.9%
	No	37212	13.1%
Baby's gender	Male	145703	51.3%
	Female	138155	48.7%
Race	Black	21080	7.4%
	White	262778	92.6%
Prenatal visit occurred	No visit	2049	0.7%
	2nd trimester	31770	11.2%
	3nd trimester	5555	2%
	Other situations	244484	86.1%
Kessner index	Adequate	225277	79.4%
	Intermediate	47557	16.7%
	Inadequate	11024	3.9%

Our outcome variable is baby's birth weights, and the treatment is mother's smoking situation, including whether smoke during pregnancy and how many cigarettes per day. In our data, 4.1% babies are low birth weight, and the mean of weights is 3454.168 grams. For the smoking information, 13% smoked during pregnancy. For the number of cigarettes smoked per day, unknown cases are given a value of 99. It is obvious that we should deal with such extreme values. The way we use to deal with it is to replace the value 99 by the average number of cigarettes a smoker would consume. Except for the outcome variable and treatment variables, there are also variables in the Natality

Data Sets that may be related with outcome, which can be used for control variables in regression. We constructed the correlation matrix of all the variables, and drew it on a heat map to choose the control variables that should be contained in regression models.

Figure 1: Heatmap of correlations



According to the heatmap, we decided to choose mom's age, education level, length of gestation, race, number of live births now living, Kessner index, marital status, and baby's gender as control variables. Although mother's year of education also have a correlation with outcome, we conclude the three binary variables generated by mother year of education instead of itself to avoid multicollinearity. Also, age square is also correlated with outcome, but it is of high value and the distribution of it is far away from normal, so we decided to only contain age in the model, which is more stable and interpretable. The Kessner index represents the adequacy of prenatal care utilization, including the start and number of prenatal visits, and it has three levels. To be rated Adequate, prenatal care must begin in the first trimester; to be rated Intermediate, care

must begin in the second trimester; and to be rated Inadequate, care must begin in the third trimester or not at all.

3 Identification strategies and regression results

3.1 Baseline model: Multivariate linear regression

We use multivariate linear regression as our baseline regression model. To estimate the causal effect, the most common way is to try to find control variables and regress the outcome by both control variables and treatment, that is, multivariate linear regression. However, this method is too simple and can't solve many endogenous problems, so we would just use it as a baseline regression model for comparison. Here is our model:

$$\boldsymbol{dbirwt = \alpha + \beta_1\, treatment + \beta_n\, control + u}$$

dbirwt means baby's birth weight (in grams), which is the outcome variable. Because there are not too many extreme values and the distribution of dbirwt is quite healthy, we don't take its naturally log form, even if the values of dbirwt are large. ***Treatment*** is the main independent variable that we are concerned about, it reflects mother's smoking behavior. The coefficient of $\boldsymbol{\beta_1}$ would be the treatment effect, and also be the causal effect we are seeking for. As observed from the heatmap, cigar and smoke are highly correlated, so we can't combine both of them in the regression model. But both of them are meaningful in researching this question. So, we decided to run two regressions with different treatments but same control variables and the outcome variable. The control variables are exactly as we mentioned above, including dmage, nlbnl, gestat, male, married, hsgrad, somecoll, collgrad, black, adeqcode2, adeqcode3, novisit, pretri2, pretri3. Here is our regression result:

Table 3: Results of linear regression

	(1)	(2)
Treatment	cigar	smoke
cigar	-13.5767	
	(0.1852)	
smoke		-222.2578
		(2.8178)
dmage	2.6344	2.5258
	(0.2113)	(0.2110)
nlbnl	43.3540	42.5421
	(0.8293)	0.8280
gestat	103.9917	103.9328
	(0.4065)	0.4059
male	139.1904	139.0326
	(1.7492)	(1.7466)
married	55.5075	45.7706
	(3.0948)	(3.1045)
hsgrad	61.5105	58.1993
	(3.5272)	(3.5236)
somecoll	92.0869	85.3493
	(3.8596)	(3.8609)
collgrad	103.6931	94.0063
	(4.0741)	(4.0819)
black	-181.3536	-181.0568
	(3.6080)	(3.5996)
adeqcode2	-65.0862	-64.4294
	(3.5683)	(3.5632)
adeqcode3	-112.4470	-112.8621
	(8.4793)	(8.4666)
novisit	-6.6933	-7.1802
	(13.2701)	(13.2504)
pretri2	32.6769	32.9926
	(4.2950)	(4.2887)
pretri3	65.5276	67.8833
	(10.4435)	(10.4283)
Constant	-906.8060	-879.0976
	(17.1042)	(17.1008)
Observations	283,858	283,858
R-squared	0.2544	0.2566

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We run our regression model with our collected data and got the estimated effect of independent variables on birthweight of babies. We found that maternal smoking was negatively associated with toddler weight, specifically, controlling for other variables.

For each child of a mother who smoked weighed approximately 222.26g less compared to a mother who did not smoke, and weighed 13.57g less for each more cigarette the mother smoked daily during pregnancy. Both of the coefficients are highly significant in 99% level.

3.2 Fixed effects model - regression on difference

A really interesting thing is that, in the given dataset, every mother had given birth to two kids. Taking advantage of this, we can obtain the causal effect through regression on the difference between weights of two kids. We found some mothers whose smoking behavior changed between the two periods of pregnancy. Controlling for other variables, the difference in weights between the first and second infant can be considered as a causal effect of smoking behavior.

We believe that the causal effect derived from this identification strategy is probably the closest to the true value. Although the dataset already contains the characteristics of multiple mothers, there are still many key variables that are not reflected in the dataset, like mother's income, weight, heredity… Those omitted variables may lead to endogenous problems. However, the omitted variable problem can be largely solved by regression on difference, because for a certain mother, most of the unobserved characteristics do not change between the two periods of pregnancy.

Different between the baseline model, some control variables became 0 for every mother after taking differences, such as marriage, black, education level…So those variables are removed from the regression model. Following are our model and regression results:

$$\Delta dbirwt = \alpha + \beta_1 \Delta treatment + \beta_n \Delta control + u$$

We get the results in the table below：

Table 4: Results of fix effects

	(1)	(2)
Treatment	cigar	smoke
cigar	-7.9843	
	(0.2881)	
smoke		-134.293
		(4.465)
dmage	2.8358	2.938
	(1.0905)	(1.090)
gestat	88.7508	88.777
	(0.5042)	(0.504)
male	142.6973	142.668
	(1.9747)	(1.974)
adeqcode2	-54.5046	-54.771
	(4.2507)	(4.249)
adeqcode3	-101.9691	-102.541
	(9.9215)	(9.916)
novisit	-7.8548	-8.272
	(16.0493)	(16.041)
pretri2	31.5509	31.874
	(5.0309)	(5.029)
pretri3	64.7515	66.452
	(12.0258)	(12.020)
Constant	67.9342	67.214
	(3.0289)	(3.027)
Observations	283,858	283,858
R-squared	0.2074	0.2082

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The regression result also shows that the causal effect of mothers' smoking to baby's weight is significantly negative. The regression analysis showed that babies born to mothers who smoked during pregnancy had a significantly lower birth weight by an average of 134.29g compared to babies born to non-smoking mothers. Furthermore, the weight of the baby was found to be reduced by 7.98g for every additional cigarette the mother smoked per day during pregnancy.

It is worth noting that the causal effect obtained using the regression on difference method is much lower than that of the baseline regression model. That may be due to the presence of too many omitted variables in the baseline regression model.

3.3 Propensity score matching

Propensity score matching is a statistical technique that creates the treatment and control groups to compare in sample data by matching individuals with similar propensity scores. In our research, for each mother who smoked during pregnancy, she would be matched up with another mother who did not smoke but had similar propensity scores. The mothers who smoked would be in the treatment group, while mothers not having smoked would be in the control group. Then we can derive the causal effect by running regression to see the difference of average infant birth weight between two groups. It is worth mentioning that in PSM, we do not use cigar as treatment because it is easier to reach a match with the binary variable smoke, as the grouping base. Matching with the variable cigar could be very difficult.

The result of PSM is that mother's smoking has a statistically significant negative effect on baby's birth weight. The coefficient for “smoke” is -223.996 with a standard error of 4.091, which indicates that babies born to smoking mothers have a birth weight that is approximately 224 grams lower than those born to non-smoking mothers on average. From the p-value and t-test, we can reject the null hypothesis that maternal smoking has no effect on baby's birth weight under 1% confidence level.

Compared to the baseline model, PSM can handle non-linear and non-additive relationships, and can reduce the bias due to observed confounders. What’s more, we have such a large dataset so that balance in the propensity score distribution is ensured.

However, there are still some limitations of this model, for we can only use the information we have to do the matching. In fact, some unobservable indices, like the personality of mothers, this paper does not capture.

3.4 Instrumental variable: dmeduc

Instrumental variable (IV) is also a statistical technique used to estimate causal effects in the presence of endogeneity, where the treatment variable is correlated with unobserved variables that also affect the outcome variable. In our research, we use the mother's education level as an instrumental variable to estimate the causal effect of smoking during pregnancy on infant birth weight. This is because education is considered to be highly correlated with smoking behavior during pregnancy but is not directly related to infant birth weight.

So, we did two IV regressions with the treatments of cigar and smoke. This time, control variables that represent education level, hsgrad, somecoll, collgrad, would no longer be in the regression model. Here is the result of the second stage regression.

Table 5: Results of IV regression

	(1)	(2)
Treatment	cigar	smoke
cigar	-33.9131	
	(0.9116)	
smoke		-475.1948
		(12.5941)
dmage	3.6717	3.1736
	(0.2043)	(0.2072)
nlbnl	44.6336	42.6659
	(0.9070)	(0.8780)
gestat	103.2548	103.3081
	(0.6060)	(0.6012)
male	138.8358	138.6004
	(1.7843)	(1.7696)
married	12.4487	2.9649
	(4.3827)	(4.4804)
black	-222.5816	-211.5949

	(4.3290)	(4.1842)
adeqcode2	-61.6802	-61.1848
	(3.7558)	(3.7174)
adeqcode3	-97.3247	-102.7594
	(9.6613)	(9.3461)
novisit	12.4154	6.2013
	(15.7103)	(15.1354)
pretri2	39.0313	38.0846
	(4.6161)	(4.5488)
pretri3	65.8020	70.7650
	(11.6765)	(11.3327)
Constant	-758.1379	-729.4895
	(25.6859)	(25.7269)
Observations	283,858	283,858
R-squared	0.2227	0.2354

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This time, the regression results showed that the causal effect was significantly negative though. However, this is much larger than the regression coefficients assessed by the other three methods. We were not sure if our instrumental variables were valid, so we performed a series of tests on the dmeduc to see if it is a good instrument variable. According to the weakening instruments test, for both treatment cigar and smoke, dmeduc are all strong instrument variable, with p value<0.01. Also, the Wu-Hausman test we conduct suggests that IV is inconsistent with OLS, indicating that there is an endogenous problem in OLS, and IV may work better than it.

Although both Wu-Hausman and weakening instruments suggests dmeduc is a good instrument variable, we still found something that may lead to the problem. That is, unlike what we thought, education may not be an appropriate exogenous variable in this problem. Education may not only affect child weight by influencing smoking behavior, but may also affect child weight by influencing income, knowledge of pregnancy preparation, and other factors, which are related to baby's weight. Therefore, the causal effect obtained with the instrumental variable regression model is likely to

be biased.

4 Discussion of finding

Through our analysis, we found that maternal smoking has a very significant effect on a baby's weight. Unfortunately, most of our strategies have endogenous problems that are difficult to solve. Among four identification strategies, we all agree that causal effects derived by regression on differences (fixed effects model) tend to be the most accurate, compared to other strategies. Since there are so many crucial variables omitted, such as a mother's weight, income, gene, and regression on difference worked so well on dealing with those unobservable factors, while other strategies did not. Actually, IV may be the most proper way to calculate the causal effect, but in our dataset we failed to find a really good instrument variable. Thus, we decided to use regression on difference as our final results, and the causal effect obtained by other strategies would be used as comparison and reference.

5 Conclusion

In this paper, we investigate the effect of maternal smoking habits on fetal weight. We first used an OLS multivariate model to study the overall trend, and then tried to use three different methods to evaluate the causal effect. They are regression on difference (fixed effects model), propensity score matching, and instrumental variable. The estimated causal effect derived from those methods are shown below:

Table 6: Results summary

Strategies	Causal effect for smoke	Causal effect for cigar
Baseline model	-222.258*** (2.817)	-13.577*** (0.185)
Regression on difference	-134.293*** (3.027)	-7.98*** (0.2881)
Propensity score matching	-223.996*** (4.091)	

Instrument variable	-475.195*** (12.594)	-33.91*** (0.912)

At last, the regression on difference method performed the best when dealing with endogenous problems among three strategies, so we decided to use the causal effect estimated by that method as our final answer.

The conclusion is that, maternal smoking has a very significant (at 1% level） negative causal effect of maternal smoking on baby's birth weight. Specifically, on average, babies born to mothers who smoked during pregnancy weighed 134.29g less than babies born to non-smoking mothers. Moreover, the weight of the baby decreased by 7.98g for each additional cigarette the mother smoked per day during pregnancy.

6 References

[1] Carter, S. Percival, T., Paterson, J. and Williams, M. (2006), 'Maternal Smoking: Risks Related to Maternal Asthma and Reduced Birth Weight in a Pacific Island Birth Cohort in New Zealand', The New Zealand Medical Journal, 119(1238).

[2] Hebal, J. R., Fox, L. N. and Sexton, M. (1988), 'Dose-Response of Birth Weight to Various Measures of Maternal Smoking During Pregnancy', Journal of Clinical Epidemiology, 41(5), pp. 483-489.

[3] Tominey, E. (2007). Maternal smoking during pregnancy and early child outcomes.

[4] Kataoka, M. C., Carvalheira, A. P. P., Ferrari, A. P., Malta, M. B., de Barros Leite Carvalhaes, M. A., & de Lima Parada, C. M. G. (2018). Smoking during pregnancy and harm reduction in birth weight: a cross-sectional study. BMC pregnancy and childbirth, 18(1), 1-10.

[5] Bogl, L. H., Strohmaier, S., Eliassen, H., Massa, J., Field, A., Chavarro, J., ... & Schernhammer, E. (2020). Maternal diet during pregnancy and child weight outcomes. The Proceedings of the Nutrition Society, 79(OCE2).

7 Appendix

```
#Load the dataset
setwd("/Users/yuqipang/Documents/570 data")
data= read.csv("birpanel.csv")

#see the summary of data
summary(data)
```

```
##      momid3            idx         stateres         dmage
   dmeduc
##  Min.   :     1   Min.   :1.0   Min.   : 1.00   Min.   :13.00   M
in.   : 0.00
##  1st Qu.: 35483   1st Qu.:1.0   1st Qu.:14.00   1st Qu.:24.00   1
st Qu.:12.00
##  Median : 70965   Median :1.5   Median :26.00   Median :28.00   M
edian :14.00
##  Mean   : 70965   Mean   :1.5   Mean   :26.61   Mean   :28.37   M
ean   :13.88
##  3rd Qu.:106447   3rd Qu.:2.0   3rd Qu.:39.00   3rd Qu.:32.00   3
rd Qu.:16.00
##  Max.   :141929   Max.   :2.0   Max.   :51.00   Max.   :50.00   M
ax.   :17.00
##      mplbir          nlbnl            gestat          dbirwt
##  Min.   : 1.00   Min.   : 0.000   Min.   :17.00   Min.   : 227
##  1st Qu.:15.00   1st Qu.: 0.000   1st Qu.:38.00   1st Qu.:3147
##  Median :26.00   Median : 1.000   Median :39.00   Median :3459
##  Mean   :26.14   Mean   : 1.188   Mean   :39.25   Mean   :3454
##  3rd Qu.:38.00   3rd Qu.: 2.000   3rd Qu.:40.00   3rd Qu.:3799
##  Max.   :51.00   Max.   :15.000   Max.   :47.00   Max.   :8020
##      cigar            smoke             male             year

##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000

##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.000

##  Median : 0.000   Median :0.0000   Median :1.0000   Median :4.000

##  Mean   : 2.131   Mean   :0.1301   Mean   :0.5133   Mean   :3.855

##  3rd Qu.: 0.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:6.000

##  Max.   :99.000   Max.   :1.0000   Max.   :1.0000   Max.   :8.000

##     married           hsgrad         somecoll         collgrad
```

```
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000

##  1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000

##  Median :1.0000   Median :0.000   Median :0.0000   Median :0.0000

##  Mean   :0.8689   Mean   :0.295   Mean   :0.2343   Mean   :0.3778

##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:1.0000

##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000

##      agesq            black           adeqcode2        adeqcode3

##  Min.   : 169.0   Min.   :0.00000   Min.   :0.0000   Min.   :0.00
000
##  1st Qu.: 576.0   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.00
000
##  Median : 784.0   Median :0.00000   Median :0.0000   Median :0.00
000
##  Mean   : 834.7   Mean   :0.07426   Mean   :0.1675   Mean   :0.03
884
##  3rd Qu.:1024.0   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.00
000
##  Max.   :2500.0   Max.   :1.00000   Max.   :1.0000   Max.   :1.00
000
##     novisit            pretri2          pretri3
##  Min.   :0.000000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.000000   Median :0.0000   Median :0.00000
##  Mean   :0.007218   Mean   :0.1119   Mean   :0.01957
##  3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:0.00000
##  Max.   :1.000000   Max.   :1.0000   Max.   :1.00000
```

```
#clean the data, convert 99(unknown) in cigar into average value.
data$cigar[data$cigar == 99] <- mean(data$cigar[data$cigar != 99 & d
ata$cigar != 0], na.rm = TRUE)

# generate and plot correlation matrix
library(ggcorrplot)

## Loading required package: ggplot2

cor_matrix <- cor(data)
cor_matrix

##                 momid3           idx      stateres         dmage
       dmeduc
## momid3     1.0000000000  0.0000000000  0.1685379755 -4.143780e-02
  0.062196329
## idx        0.0000000000  1.0000000000  0.0000000000  2.266642e-01
  0.000000000
## stateres   0.1685379755  0.0000000000  1.0000000000  7.858944e-03
  0.011068053
```

```
## dmage     -0.0414378026  0.2266642361  0.0078589444  1.000000e+00
  0.510036103
## dmeduc     0.0621963288  0.0000000000  0.0110680530  5.100361e-01
  1.000000000
## mplbir    -0.0057765780  0.0000000000 -0.0312839587  2.745268e-02
  0.022844062
## nlbnl     -0.0771635516  0.4155796786  0.0248736245  3.031937e-01
 -0.139084778
## gestat    -0.0085947931 -0.0413269171  0.0024638620 -1.643448e-02
  0.020703013
## dbirwt    -0.0058699817  0.0523189205  0.0116913429  1.198694e-01
  0.132178618
## cigar     -0.0452688474  0.0145988992 -0.0035374984 -1.249437e-01
 -0.292522457
## smoke     -0.0404895893  0.0077279008  0.0004397186 -1.630107e-01
 -0.328264906
## male      -0.0005732736 -0.0058571113  0.0020687410  8.751667e-05
  0.005356821
## year       0.7232944217  0.5459217813  0.0134302529  1.515114e-01
  0.058300137
## married   -0.0009017620  0.0000000000  0.0170557501  3.141455e-01
  0.373919298
## hsgrad    -0.0552795729  0.0000000000 -0.0089662932 -2.343755e-01
 -0.536191478
## somecoll   0.0010193538  0.0000000000  0.0025426828  2.376301e-03
 -0.015407334
## collgrad   0.0598671413  0.0000000000  0.0087139600  4.262056e-01
  0.852596130
## agesq     -0.0387767153  0.2274137574  0.0066425065  9.933940e-01
  0.484132279
## black     -0.0191359018  0.0000000000 -0.0398433731 -1.264517e-01
 -0.146043646
## adeqcode2 -0.0192848552  0.0096124519  0.0171217998 -1.209812e-01
 -0.168302338
## adeqcode3 -0.0187623559  0.0063089524 -0.0037078111 -9.835245e-02
 -0.176687855
## novisit   -0.0127175949  0.0049522043 -0.0060105785 -2.840090e-02
 -0.076364816
## pretri2   -0.0252105865  0.0011621138 -0.0015105672 -1.491352e-01
 -0.200847709
## pretri3   -0.0121761082  0.0008901579 -0.0037370358 -8.319470e-02
 -0.128045545
##                  mplbir        nlbnl       gestat       dbirwt
     cigar
## momid3    -5.776578e-03 -0.077163552 -0.008594793 -0.005869982 -
0.045268847
## idx        0.000000e+00  0.415579679 -0.041326917  0.052318921
0.014598899
## stateres  -3.128396e-02  0.024873624  0.002463862  0.011691343 -
0.003537498
## dmage      2.745268e-02  0.303193699 -0.016434483  0.119869413 -
0.124943693
## dmeduc     2.284406e-02 -0.139084778  0.020703013  0.132178618 -
0.292522457
```

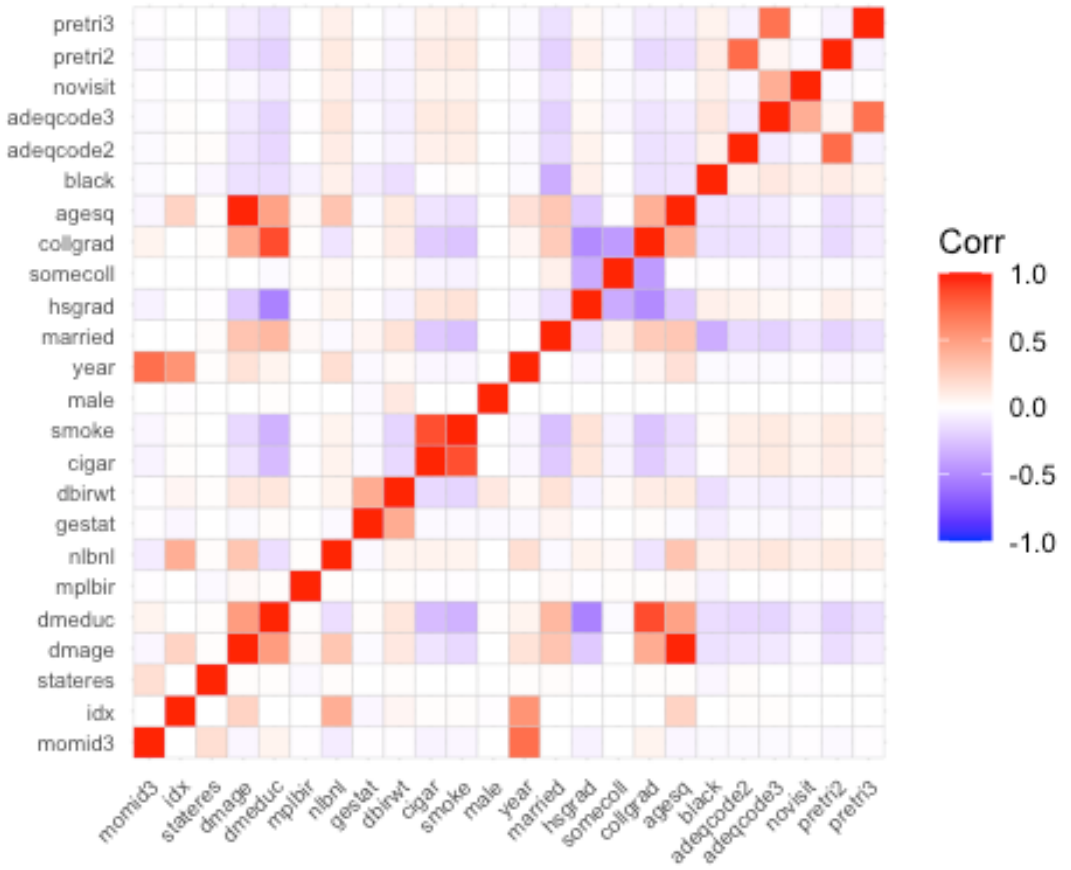
```
## mplbir     1.000000e+00  0.008768059  0.002811841  0.010126925 -
0.011288043
## nlbnl      8.768059e-03  1.000000000 -0.028942393  0.062091647
0.072404743
## gestat     2.811841e-03 -0.028942393  1.000000000  0.427045989 -
0.025112050
## dbirwt     1.012692e-02  0.062091647  0.427045989  1.000000000 -
0.163929172
## cigar     -1.128804e-02  0.072404743 -0.025112050 -0.163929172
1.000000000
## smoke     -1.257265e-02  0.058589786 -0.027570182 -0.181663032
0.836631398
## male       1.983705e-03 -0.004188590 -0.028911895  0.117612637 -
0.002783618
## year       7.417111e-05  0.170847369 -0.031402706  0.029767506 -
0.035482995
## married    3.107602e-02 -0.033284269  0.052970258  0.154827915 -
0.226864864
## hsgrad    -7.986018e-03  0.056934258 -0.007241539 -0.055973587
0.127142571
## somecoll   1.272835e-02  0.025440965  0.002405393  0.028945038 -
0.054213368
## collgrad   1.227459e-02 -0.123353560  0.017526925  0.098076553 -
0.217905072
## agesq      2.625091e-02  0.307821029 -0.019754368  0.112874738 -
0.118960451
## black     -5.742496e-02  0.075843677 -0.081427076 -0.144466859 -
0.007106225
## adeqcode2 -3.982230e-03  0.099769992 -0.021690596 -0.063519870
0.083141812
## adeqcode3 -7.379060e-03  0.126207935 -0.034378453 -0.071392592
0.106711265
## novisit   -3.876281e-03  0.083610061 -0.046303712 -0.052495261
0.060569602
## pretri2   -7.428303e-03  0.107261453  0.009179432 -0.045883338
0.100822758
## pretri3   -3.197212e-03  0.075660891  0.001876181 -0.031202684
0.069174058
##                   smoke          male          year      married
       hsgrad
## momid3    -0.0404895893 -5.732736e-04  7.232944e-01 -0.000901762
-0.0552795729
## idx        0.0077279008 -5.857111e-03  5.459218e-01  0.000000000
 0.0000000000
## stateres   0.0004397186  2.068741e-03  1.343025e-02  0.017055750
-0.0089662932
## dmage     -0.1630107138  8.751667e-05  1.515114e-01  0.314145545
-0.2343755310
## dmeduc    -0.3282649056  5.356821e-03  5.830014e-02  0.373919298
-0.5361914783
## mplbir    -0.0125726476  1.983705e-03  7.417111e-05  0.031076021
-0.0079860176
## nlbnl      0.0585897859 -4.188590e-03  1.708474e-01 -0.033284269
 0.0569342578
```

```
## gestat    -0.0275701820 -2.891189e-02 -3.140271e-02  0.052970258
-0.0072415393
## dbirwt    -0.1816630321  1.176126e-01  2.976751e-02  0.154827915
-0.0559735867
## cigar      0.8366313980 -2.783618e-03 -3.548300e-02 -0.226864864
 0.1271425712
## smoke      1.0000000000 -3.814022e-03 -3.561600e-02 -0.272291149
 0.1492747294
## male      -0.0038140219  1.000000e+00 -4.009013e-03  0.002271041
-0.0005297428
## year      -0.0356159959 -4.009013e-03  1.000000e+00  0.002148954
-0.0413877561
## married   -0.2722911489  2.271041e-03  2.148954e-03  1.000000000
-0.1417042276
## hsgrad     0.1492747294 -5.297428e-04 -4.138776e-02 -0.141704228
 1.0000000000
## somecoll  -0.0572526271  8.256246e-04  4.465203e-03  0.075309466
-0.3578500578
## collgrad  -0.2503581198  3.333226e-03  5.122755e-02  0.274837657
-0.5040141471
## agesq     -0.1544476278  1.881786e-04  1.553037e-01  0.287337972
-0.2302771913
## black      0.0175000148 -4.792077e-03 -1.724104e-02 -0.349315107
 0.0831347258
## adeqcode2  0.0942441551  4.439176e-03 -2.546052e-02 -0.155278631
 0.0727577272
## adeqcode3  0.1149344419  7.088857e-04 -2.395042e-02 -0.199887060
 0.0431054933
## novisit    0.0614024794  3.545006e-04 -1.242435e-02 -0.108308131
 0.0221359247
## pretri2    0.1129365826  4.842492e-03 -3.541739e-02 -0.191239043
 0.0841992599
## pretri3    0.0778812853 -2.725272e-04 -1.931181e-02 -0.134569988
 0.0303586333
##                somecoll     collgrad         agesq        black
  adeqcode2
## momid3     0.0010193538  0.059867141 -0.0387767153 -0.019135902 -
0.019284855
## idx        0.0000000000  0.000000000  0.2274137574  0.000000000
0.009612452
## stateres   0.0025426828  0.008713960  0.0066425065 -0.039843373
0.017121800
## dmage      0.0023763009  0.426205582  0.9933940054 -0.126451746 -
0.120981239
## dmeduc    -0.0154073336  0.852596130  0.4841322787 -0.146043646 -
0.168302338
## mplbir     0.0127283481  0.012274590  0.0262509144 -0.057424959 -
0.003982230
## nlbnl      0.0254409646 -0.123353560  0.3078210295  0.075843677
0.099769992
## gestat     0.0024053934  0.017526925 -0.0197543684 -0.081427076 -
0.021690596
## dbirwt     0.0289450383  0.098076553  0.1128747382 -0.144466859 -
0.063519870
```

```
## cigar     -0.0542133683 -0.217905072 -0.1189604510 -0.007106225
0.083141812
## smoke     -0.0572526271 -0.250358120 -0.1544476278  0.017500015
0.094244155
## male       0.0008256246  0.003333226  0.0001881786 -0.004792077
0.004439176
## year       0.0044652027  0.051227546  0.1553037382 -0.017241040 -
0.025460519
## married    0.0753094656  0.274837657  0.2873379719 -0.349315107 -
0.155278631
## hsgrad    -0.3578500578 -0.504014147 -0.2302771913  0.083134726
0.072757727
## somecoll   1.0000000000 -0.431063405 -0.0099987828 -0.010524407 -
0.013257205
## collgrad  -0.4310634050  1.000000000  0.4125010670 -0.131066610 -
0.134558308
## agesq     -0.0099987828  0.412501067  1.0000000000 -0.117540133 -
0.109852284
## black     -0.0105244073 -0.131066610 -0.1175401332  1.000000000
0.082687121
## adeqcode2 -0.0132572053 -0.134558308 -0.1098522844  0.082687121
1.000000000
## adeqcode3 -0.0379379350 -0.115593879 -0.0886338252  0.124782849 -
0.090176796
## novisit   -0.0198606905 -0.050904779 -0.0245480425  0.078221385 -
0.038253220
## pretri2   -0.0203348141 -0.157273477 -0.1361269054  0.102907348
0.730800581
## pretri3   -0.0241211980 -0.084379810 -0.0753827342  0.068042559 -
0.063380773
##               adeqcode3       novisit      pretri2       pretri3
## momid3    -0.0187623559 -0.0127175949 -0.025210587 -0.0121761082
## idx        0.0063089524  0.0049522043  0.001162114  0.0008901579
## stateres  -0.0037078111 -0.0060105785 -0.001510567 -0.0037370358
## dmage     -0.0983524504 -0.0284008964 -0.149135203 -0.0831946960
## dmeduc    -0.1766878545 -0.0763648162 -0.200847709 -0.1280455455
## mplbir    -0.0073790601 -0.0038762814 -0.007428303 -0.0031972117
## nlbnl      0.1262079346  0.0836100615  0.107261453  0.0756608906
## gestat    -0.0343784530 -0.0463037118  0.009179432  0.0018761813
## dbirwt    -0.0713925917 -0.0524952612 -0.045883338 -0.0312026845
## cigar      0.1067112653  0.0605696025  0.100822758  0.0691740577
## smoke      0.1149344419  0.0614024794  0.112936583  0.0778812853
## male       0.0007088857  0.0003545006  0.004842492 -0.0002725272
## year      -0.0239504233 -0.0124243461 -0.035417388 -0.0193118141
## married   -0.1998870598 -0.1083081311 -0.191239043 -0.1345699879
## hsgrad     0.0431054933  0.0221359247  0.084199260  0.0303586333
## somecoll  -0.0379379350 -0.0198606905 -0.020334814 -0.0241211980
## collgrad  -0.1155938792 -0.0509047789 -0.157273477 -0.0843798097
## agesq     -0.0886338252 -0.0245480425 -0.136126905 -0.0753827342
## black      0.1247828491  0.0782213848  0.102907348  0.0680425592
## adeqcode2 -0.0901767964 -0.0382532197  0.730800581 -0.0633807732
## adeqcode3  1.0000000000  0.4242024697  0.045642375  0.7028501317
## novisit    0.4242024697  1.0000000000 -0.030270951 -0.0120469370
```

```
## pretri2    0.0456423750 -0.0302709514  1.000000000 -0.0501551587
## pretri3    0.7028501317 -0.0120469370 -0.050155159  1.0000000000
```

```
ggcorrplot(cor_matrix,type=c("full"), tl.cex=7)
```




```
#choose control variables based on correlation matrix
control_vars <- c("dmage", "nlbnl", "gestat", "male", "married", "hs
grad", "somecoll", "collgrad", "black", "adeqcode2", "adeqcode3", "n
ovisit", "pretri2", "pretri3")
# dmeduc is not chosen. Cause hsgrad, somecoll, collgrad are generat
ed by demeduc. agesq is also not used, since it has abnormal distrib
ution and high variance, given we already had age as control varibl
e.
```

#for each method, run twice using different treatment(smoke or cigar), except PSM

#Method1:Baseline regression(multiple linear regression)

```
#Baseline model using smoke as treatment
baseline_smoke <- lm(dbirwt ~ smoke + dmage + nlbnl + gestat + male
+ married + hsgrad + somecoll + collgrad + black + adeqcode2 + adeqc
ode3 + novisit + pretri2 + pretri3, data = data)
summary(baseline_smoke)
```

```
##
## Call:
## lm(formula = dbirwt ~ smoke + dmage + nlbnl + gestat + male +
##     married + hsgrad + somecoll + collgrad + black + adeqcode2 +
```

```
##     adeqcode3 + novisit + pretri2 + pretri3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3022.2  -298.5    -6.4   293.3  4832.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -879.0976    17.1008 -51.407  < 2e-16 ***
## smoke       -222.2578     2.8178 -78.875  < 2e-16 ***
## dmage          2.5258     0.2110  11.973  < 2e-16 ***
## nlbnl         42.5421     0.8280  51.379  < 2e-16 ***
## gestat       103.9328     0.4059 256.063  < 2e-16 ***
## male         139.0326     1.7466  79.602  < 2e-16 ***
## married       45.7706     3.1045  14.743  < 2e-16 ***
## hsgrad        58.1993     3.5236  16.517  < 2e-16 ***
## somecoll      85.3493     3.8609  22.106  < 2e-16 ***
## collgrad      94.0063     4.0819  23.030  < 2e-16 ***
## black       -181.0568     3.5996 -50.299  < 2e-16 ***
## adeqcode2    -64.4294     3.5632 -18.082  < 2e-16 ***
## adeqcode3   -112.8621     8.4666 -13.330  < 2e-16 ***
## novisit       -7.1802    13.2504  -0.542    0.588
## pretri2       32.9926     4.2887   7.693 1.44e-14 ***
## pretri3       67.8833    10.4283   6.510 7.55e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 464.9 on 283842 degrees of freedom
## Multiple R-squared:  0.2566, Adjusted R-squared:  0.2565
## F-statistic:  6531 on 15 and 283842 DF,  p-value: < 2.2e-16

coef_smoke <- summary(baseline_smoke)$coefficients["smoke", ]
tvalue_smoke <- coef_smoke["t value"]
pvalue_smoke <- coef_smoke["Pr(>|t|)"]

cat("Coefficient for smoke:", coef_smoke[1], "\n")

## Coefficient for smoke: -222.2578

cat("t-value for smoke:", tvalue_smoke, "\n")

## t-value for smoke: -78.87533

cat("p-value for smoke:", pvalue_smoke, "\n")

## p-value for smoke: 0

#Baseline model using cigar as treatment
baseline_cigar <- lm(dbirwt ~ cigar + dmage + nlbnl + gestat + male
+ married + hsgrad + somecoll + collgrad + black + adeqcode2 + adeqc
ode3 + novisit + pretri2 + pretri3, data = data)
summary(baseline_cigar)

##
## Call:
```

```
## lm(formula = dbirwt ~ cigar + dmage + nlbnl + gestat + male +
##     married + hsgrad + somecoll + collgrad + black + adeqcode2 +
##     adeqcode3 + novisit + pretri2 + pretri3, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3022.8  -298.8    -6.0   293.8  4852.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -906.8060    17.1042 -53.017  < 2e-16 ***
## cigar        -13.5767     0.1852 -73.299  < 2e-16 ***
## dmage          2.6344     0.2113  12.465  < 2e-16 ***
## nlbnl         43.3540     0.8293  52.276  < 2e-16 ***
## gestat       103.9917     0.4065 255.836  < 2e-16 ***
## male         139.1904     1.7492  79.576  < 2e-16 ***
## married       55.5075     3.0948  17.936  < 2e-16 ***
## hsgrad        61.5105     3.5272  17.439  < 2e-16 ***
## somecoll      92.0869     3.8596  23.859  < 2e-16 ***
## collgrad     103.6931     4.0741  25.452  < 2e-16 ***
## black       -181.3536     3.6080 -50.264  < 2e-16 ***
## adeqcode2    -65.0862     3.5683 -18.240  < 2e-16 ***
## adeqcode3   -112.4470     8.4793 -13.261  < 2e-16 ***
## novisit       -6.3933    13.2701  -0.482     0.63
## pretri2       32.6769     4.2950   7.608 2.79e-14 ***
## pretri3       65.5276    10.4435   6.274 3.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 465.6 on 283842 degrees of freedom
## Multiple R-squared:  0.2544, Adjusted R-squared:  0.2544
## F-statistic:  6457 on 15 and 283842 DF,  p-value: < 2.2e-16

coef_cigar <- summary(baseline_cigar)$coefficients["cigar", ]
tvalue_cigar <- coef_cigar["t value"]
pvalue_cigar <- coef_cigar["Pr(>|t|)"]

# Print the results
cat("Coefficient for cigar:", coef_cigar[1], "\n")

## Coefficient for cigar: -13.57675

cat("t-value for cigar:", tvalue_cigar, "\n")

## t-value for cigar: -73.29927

cat("p-value for cigar:", pvalue_cigar, "\n")

## p-value for cigar: 0
```

#Method2:Fixed effects model: regression on difference

```
# generate difference between two pregnancy for variables
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

diff = data %>%
  group_by(momid3) %>%
  mutate(idx = idx - lag(idx)) %>%
  mutate(stateres = stateres - lag(stateres)) %>%
  mutate(dmage = dmage - lag(dmage)) %>%
  mutate(dmeduc = dmeduc - lag(dmeduc)) %>%
  mutate(mplbir = mplbir - lag(mplbir)) %>%
  mutate(nlbnl = nlbnl - lag(nlbnl)) %>%
  mutate(gestat = gestat - lag(gestat)) %>%
  mutate(dbirwt = dbirwt - lag(dbirwt)) %>%
  mutate(cigar = cigar - lag(cigar)) %>%
  mutate(smoke = smoke - lag(smoke)) %>%
  mutate(male = male - lag(male)) %>%
  mutate(year = year - lag(year)) %>%
  mutate(married = married - lag(married)) %>%
  mutate(hsgrad = hsgrad - lag(hsgrad)) %>%
  mutate(somecoll = somecoll - lag(somecoll)) %>%
  mutate(collgrad = collgrad - lag(collgrad)) %>%
  mutate(agesq = agesq - lag(agesq)) %>%
  mutate(black = black - lag(black)) %>%
  mutate(adeqcode2 = adeqcode2 - lag(adeqcode2)) %>%
  mutate(adeqcode3 = adeqcode3 - lag(adeqcode3)) %>%
  mutate(novisit = novisit - lag(novisit)) %>%
  mutate(pretri2 = pretri2 - lag(pretri2)) %>%
  mutate(pretri3 = pretri3 - lag(pretri3)) %>%
  slice_tail(n = 1)

summary(diff)

##      momid3            idx       stateres     dmage           dme
duc
##  Min.   :     1   Min.   :1   Min.   :0   Min.   :0.000   Min.
 :0
##  1st Qu.: 35483   1st Qu.:1   1st Qu.:0   1st Qu.:2.000   1st Q
u.:0
##  Median : 70965   Median :1   Median :0   Median :2.000   Median
 :0
##  Mean   : 70965   Mean   :1   Mean   :0   Mean   :2.464   Mean
 :0
##  3rd Qu.:106447   3rd Qu.:1   3rd Qu.:0   3rd Qu.:3.000   3rd Q
u.:0
##  Max.   :141929   Max.   :1   Max.   :0   Max.   :9.000   Max.
 :0
##      mplbir      nlbnl       gestat            dbirwt
```

```
##  Min.   :0   Min.   :1   Min.   :-24.000   Min.   :-4628.00
##  1st Qu.:0   1st Qu.:1   1st Qu.: -2.000   1st Qu.: -284.00
##  Median :0   Median :1   Median :  0.000   Median :   57.00
##  Mean   :0   Mean   :1   Mean   : -0.179   Mean   :   56.42
##  3rd Qu.:0   3rd Qu.:1   3rd Qu.:  1.000   3rd Qu.:  397.00
##  Max.   :0   Max.   :1   Max.   : 22.000   Max.   : 5327.00
##      cigar              smoke              male                ye
ar
##  Min.   :-98.0000   Min.   :-1.0000   Min.   :-1.000000   Min.
 :0.000
##  1st Qu.:  0.0000   1st Qu.: 0.0000   1st Qu.:-1.000000   1st Q
u.:2.000
##  Median :  0.0000   Median : 0.0000   Median : 0.000000   Median
 :2.000
##  Mean   :  0.1471   Mean   : 0.0052   Mean   :-0.005855   Mean
 :2.467
##  3rd Qu.:  0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.000000   3rd Q
u.:3.000
##  Max.   : 93.0000   Max.   : 1.0000   Max.   : 1.000000   Max.
 :8.000
##     married      hsgrad     somecoll    collgrad     agesq
    black
##  Min.   :0   Min.   :0   Min.   :0   Min.   :0   Min.   :  0.0
Min.   :0
##  1st Qu.:0   1st Qu.:0   1st Qu.:0   1st Qu.:0   1st Qu.: 84.0
1st Qu.:0
##  Median :0   Median :0   Median :0   Median :0   Median :124.0
Median :0
##  Mean   :0   Mean   :0   Mean   :0   Mean   :0   Mean   :141.5
Mean   :0
##  3rd Qu.:0   3rd Qu.:0   3rd Qu.:0   3rd Qu.:0   3rd Qu.:183.0
3rd Qu.:0
##  Max.   :0   Max.   :0   Max.   :0   Max.   :0   Max.   :672.0
Max.   :0
##    adeqcode2          adeqcode3            novisit
##  Min.   :-1.00000   Min.   :-1.000000   Min.   :-1.0000000
##  1st Qu.: 0.00000   1st Qu.: 0.000000   1st Qu.: 0.0000000
##  Median : 0.00000   Median : 0.000000   Median : 0.0000000
##  Mean   : 0.00718   Mean   : 0.002438   Mean   : 0.0008384
##  3rd Qu.: 0.00000   3rd Qu.: 0.000000   3rd Qu.: 0.0000000
##  Max.   : 1.00000   Max.   : 1.000000   Max.   : 1.0000000
##     pretri2              pretri3
##  Min.   :-1.0000000   Min.   :-1.0000000
##  1st Qu.: 0.0000000   1st Qu.: 0.0000000
##  Median : 0.0000000   Median : 0.0000000
##  Mean   : 0.0007328   Mean   : 0.0002466
##  3rd Qu.: 0.0000000   3rd Qu.: 0.0000000
##  Max.   : 1.0000000   Max.   : 1.0000000
```

```
# regression on difference. Treatment=smoke
diff_smoke = lm(dbirwt ~ smoke + dmage + gestat + male  + adeqcode2
+ adeqcode3 + novisit + pretri2 + pretri3, data = diff)
summary(diff_smoke)
```

```
##
## Call:
## lm(formula = dbirwt ~ smoke + dmage + gestat + male + adeqcode2 +

##     adeqcode3 + novisit + pretri2 + pretri3, data = diff)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4580.4  -321.2     0.7   321.5  5288.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.214      3.027  22.204  < 2e-16 ***
## smoke       -134.293      4.465 -30.077  < 2e-16 ***
## dmage          2.938      1.090   2.695  0.00704 **
## gestat        88.777      0.504 176.158  < 2e-16 ***
## male         142.668      1.974  72.282  < 2e-16 ***
## adeqcode2    -54.771      4.249 -12.892  < 2e-16 ***
## adeqcode3   -102.541      9.916 -10.340  < 2e-16 ***
## novisit       -8.272     16.041  -0.516  0.60608
## pretri2       31.874      5.029   6.339 2.32e-10 ***
## pretri3       66.452     12.020   5.528 3.24e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 525 on 141919 degrees of freedom
## Multiple R-squared:  0.2082, Adjusted R-squared:  0.2081
## F-statistic:  4146 on 9 and 141919 DF,  p-value: < 2.2e-16

coef_smoke <- summary(diff_smoke)$coefficients["smoke", ]
tvalue_smoke <- coef_smoke["t value"]
pvalue_smoke <- coef_smoke["Pr(>|t|)"]

cat("Coefficient for smoke:", coef_smoke[1], "\n")

## Coefficient for smoke: -134.2926

cat("t-value for smoke:", tvalue_smoke, "\n")

## t-value for smoke: -30.07747

cat("p-value for smoke:", pvalue_smoke, "\n")

## p-value for smoke: 4.026753e-198

# regression on difference. Treatment=smoke
diff_cigar = lm(dbirwt ~ cigar + dmage + gestat + male  + adeqcode2
+ adeqcode3 + novisit + pretri2 + pretri3, data = diff)
summary(diff_cigar)

##
## Call:
## lm(formula = dbirwt ~ cigar + dmage + gestat + male + adeqcode2 +

##     adeqcode3 + novisit + pretri2 + pretri3, data = diff)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4580.9  -321.0     1.3   321.3  5342.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   67.9342     3.0289  22.429  < 2e-16 ***
## cigar         -7.9843     0.2881 -27.714  < 2e-16 ***
## dmage          2.8358     1.0905   2.600  0.00931 **
## gestat        88.7508     0.5042 176.018  < 2e-16 ***
## male         142.6973     1.9747  72.262  < 2e-16 ***
## adeqcode2    -54.5046     4.2507 -12.822  < 2e-16 ***
## adeqcode3   -101.9691     9.9215 -10.278  < 2e-16 ***
## novisit       -7.8548    16.0493  -0.489  0.62455
## pretri2       31.5509     5.0309   6.271 3.59e-10 ***
## pretri3       64.7515    12.0258   5.384 7.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 525.3 on 141919 degrees of freedom
## Multiple R-squared:  0.2074, Adjusted R-squared:  0.2074
## F-statistic:  4127 on 9 and 141919 DF,  p-value: < 2.2e-16

coef_cigar <- summary(diff_cigar)$coefficients["cigar", ]
tvalue_cigar <- coef_smoke["t value"]
pvalue_cigar <- coef_smoke["Pr(>|t|)"]

cat("Coefficient for cigar:", coef_cigar[1], "\n")

## Coefficient for cigar: -7.984308

cat("t-value for cigar:", tvalue_cigar, "\n")

## t-value for cigar: -30.07747

cat("p-value for cigar:", pvalue_cigar, "\n")

## p-value for cigar: 4.026753e-198
```

#Method3:PSM Propensity Score Matching

```
library(dplyr)
library(MatchIt)
vars_to_match <- c("dmage", "nlbnl", "gestat", "male", "married", "h
sgrad", "somecoll", "collgrad", "black", "adeqcode2", "adeqcode3", "
novisit", "pretri2", "pretri3")
data_match <- data %>% dplyr::select(dbirwt, smoke, one_of(vars_to_m
atch))
m.out <- matchit(smoke ~ dmage + nlbnl + gestat + male + married + h
sgrad + somecoll + collgrad + black + adeqcode2 + adeqcode3 + novisi
t + pretri2 + pretri3, data = data_match, method = "nearest")
matched_data <- match.data(m.out)
PSM_smoke<- lm(dbirwt ~ smoke, data = matched_data)
summary(PSM_smoke)
```

```
##
## Call:
## lm(formula = dbirwt ~ smoke, data = matched_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3197.9  -308.9    20.1   346.1  4595.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3424.915      2.893 1184.05   <2e-16 ***
## smoke       -223.996      4.091  -54.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 555.9 on 73866 degrees of freedom
## Multiple R-squared:  0.03901,    Adjusted R-squared:  0.039
## F-statistic:  2998 on 1 and 73866 DF,  p-value: < 2.2e-16
```

#Method4:Instrument variable: dmeduc

```
# Treatment=smoke
library(AER)

## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

iv_smoke <- ivreg(dbirwt ~ smoke + dmage + nlbnl + gestat + male + m
arried + black + adeqcode2 + adeqcode3 + novisit + pretri2 + pretri3
 | dmeduc+ dmage + nlbnl + gestat + male + married + black + adeqcod
e2 + adeqcode3 + novisit + pretri2 + pretri3, data = data)

summary(iv_smoke, vcov = sandwich, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = dbirwt ~ smoke + dmage + nlbnl + gestat + male +
##     married + black + adeqcode2 + adeqcode3 + novisit + pretri2 +

##     pretri3 | dmeduc + dmage + nlbnl + gestat + male + married +
##     black + adeqcode2 + adeqcode3 + novisit + pretri2 + pretri3,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -3047.816  -303.398    -7.266   297.779  4743.916
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -729.4895    25.7269 -28.355  < 2e-16 ***
## smoke       -475.1948    12.5941 -37.732  < 2e-16 ***
## dmage          3.1736     0.2072  15.318  < 2e-16 ***
## nlbnl         42.6659     0.8780  48.596  < 2e-16 ***
## gestat       103.3081     0.6012 171.844  < 2e-16 ***
## male         138.6004     1.7696  78.323  < 2e-16 ***
## married        2.9649     4.4804   0.662    0.508
## black       -211.5949     4.1842 -50.570  < 2e-16 ***
## adeqcode2    -61.1848     3.7174 -16.459  < 2e-16 ***
## adeqcode3   -102.7594     9.3461 -10.995  < 2e-16 ***
## novisit        6.2013    15.1354   0.410    0.682
## pretri2       38.0846     4.5488   8.372  < 2e-16 ***
## pretri3       70.7650    11.3327   6.244 4.26e-10 ***
##
## Diagnostic tests:
##                      df1    df2 statistic p-value
## Weak instruments       1 283845   13070.4  <2e-16 ***
## Wu-Hausman             1 283844     381.6  <2e-16 ***
## Sargan                 0     NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.5 on 283845 degrees of freedom
## Multiple R-Squared: 0.2354,  Adjusted R-squared: 0.2353
## Wald test:  4292 on 12 and 283845 DF,  p-value: < 2.2e-16

# Treatment=cigar
iv_cigar <- ivreg(dbirwt ~ cigar + dmage + nlbnl + gestat + male + m
arried + black + adeqcode2 + adeqcode3 + novisit + pretri2 + pretri3
 | dmeduc+ dmage + nlbnl + gestat + male + married + black + adeqcod
e2 + adeqcode3 + novisit + pretri2 + pretri3, data = data)

summary(iv_cigar, vcov = sandwich, diagnostics = TRUE)

##

##                                                              Call:

## ivreg(formula = dbirwt ~ cigar + dmage + nlbnl + gestat + male +
```

```
##     married + black + adeqcode2 + adeqcode3 + novisit + pretri2 +
##     pretri3 | dmeduc + dmage + nlbnl + gestat + male + married +
##     black + adeqcode2 + adeqcode3 + novisit + pretri2 + pretri3,
##                                         data        =        data)
##
##                                                          Residuals:
##          Min            1Q      Median            3Q           Max
##   -3056.432     -305.725          -9.338       297.523     4762.710
##
##                                                       Coefficients:
##                           Estimate  Std.  Error  t  value  Pr(>|t|)
##  (Intercept)  -758.1379         25.6859  -29.516    <  2e-16  ***
##  cigar              -33.9131         0.9116  -37.201    <  2e-16  ***
##  dmage                3.6717         0.2043   17.969    <  2e-16  ***
##  nlbnl               44.6336         0.9070   49.210    <  2e-16  ***
##  gestat             103.2548         0.6060  170.401    <  2e-16  ***
##  male               138.8358         1.7843   77.811    <  2e-16  ***
##  married             12.4487         4.3827     2.840    0.00451  **
##  black             -222.5816         4.3290  -51.417    <  2e-16  ***
##  adeqcode2         -61.6802         3.7558  -16.423    <  2e-16  ***
##  adeqcode3         -97.3247         9.6613  -10.074    <  2e-16  ***
##  novisit              12.4154        15.7103      0.790    0.42937
##  pretri2             39.0313         4.6161     8.456    <  2e-16  ***
##  pretri3             65.8020        11.6765     5.635  1.75e-08  ***
##
##                              Diagnostic                          tests:
##                                    df1       df2  statistic  p-value
```

```
##  Weak  instruments            1  283845        9775.6    <2e-16  ***
##  Wu-Hausman                    1  283844         487.1    <2e-16  ***
##  Sargan                        0        NA            NA         NA
##                                                                   ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Residual standard error: 475.4 on 283845 degrees of freedom
##   Multiple   R-Squared:   0.2227,     Adjusted   R-squared:   0.2226
## Wald test:  4220 on 12 and 283845 DF,  p-value: < 2.2e-16
```