The Socioeconomic Status of Pregnant Women with Serious Mental Illness in the United States from 2016 to 2019

Yao Jiang, Qin Li, Hantong Hu, Ivy Zhang

2021-12-16

Introduction

Among pregnant women, mental illness is common, within which depression and anxiety are the most common. The prevalence of depression and anxiety during pregnancy is about 12.4% and 13% (Cook et al., 2010), respectively, in Switzerland. Prior research has investigated the relationship between economic status and serious mental illness among parents (Luciano, Nicholson & Meara, 2014) and the prevalence of suicidal behaviors during pregnancy (Kitsantas,2020). It leads us to become curious about how income level and employment status as socioeconomic status factors are associated with the serious mental illness of pregnant women. Understanding the specific factors associated with serious mental illness could help inform the healthcare sector on better delivering mental health care to targeted populations.

Though there were fluctuations of mental illness prevalence each year, we believe that the general trend of mental illness is consistent from 2016 to 2019 for two reasons. The previous research indicated no differences in the trends in mental health among pregnant women in the United States from 2008 to 2014 (Salameh et al., 2020). In addition, 2016 to 2019 were under the Trump Administration and also before the onset of the Covid pandemic, meaning the overall political and healthcare environment were relatively stable over these four years.

We also believe that pregnant women's education level, race, age, health status, substance use order (Luciano, Nicholson & Meara, 2014), and marital status (Kitsantas,2020) may be associated with both socioeconomic status and serious mental illness status. In this study, we will adjust these potential confounders to further study the association between socioeconomic status and serious mental illness status.

Therefore, this study will address the following questions: (1) What is the prevalence of serious mental illness among pregnant women from 2016 to 2019? (2) After adjusting for the potential confounders, how do pregnant women with serious mental illnesses differ from those without serious mental illness in family income or employment status?

Data

In this study, we used data from the National Survey on Drug Use and Health (NSDUH) from 2016 to 2019. The NSDUH is a nationally representative cross-sectional survey conducted annually by the Substance Abuse and Mental Health Services Administration (SAMHSA) to assess the prevalence and frequency of the use of illicit drugs, alcohol, and tobacco and mental health issues in the USA civilians aged 12 years and older. From 2016 to 2019, the survey randomly sampled 225622 participants in total, including 2866 pregnant women aged 18 years or older. We used the data of 2754 out of 2866 pregnant women who have available data in all our interested variables to conduct further analysis.

Mental health was measured in multiple ways in the data, but we would only focus on serious mental illness (SMI) in this analysis. SMI is measured using the Kessler-6 (K6) distress scale. K6 contains a series of six questions, asking adult respondents how frequently they experienced symptoms of psychological distress during the past 30 days. K6 scores range from 0 to 24, indicating the level of psychological distress with higher scores showing higher distress level. Generally speaking, a participant with a K6 score that is 13 or higher will be defined as having SMI (Prochaska et al., 2012). We will use an indicator variable of whether the participant has a K6 score of 13 or higher as our primary outcome variable.

NSDUH recorded the family income and employment status into multi-categorical variables. In this analysis, we decided to dichotomize both these multi-categorical variables into binary variables to avoid future complexity. Participants with a family annual income of fewer than 50,000 dollars were defined as low-income, and otherwise high-income. For employment status, we would define participants who reported them as unemployed, disabled, or in school as unemployed, and otherwise employed. The binary income indicator variable and employment status variable will be our main interested dependent variables in this analysis.

The demographic characteristics assessed in this study were race, education, marital status, health status, education, and past month illicit drug use. Illicit drug use was defined as individuals who used marijuana, hallucinogens, inhalants, methamphetamine, tranquilizers, cocaine, heroin, pain relievers, stimulants, and/or sedatives. These variables, as stated in the introduction, would be the confounders in this study. Since these variables were not our main interest variables, we would not introduce them in depth. Details about how these variables are defined and coded can be found on the website of NSDUH.

Methods

The first aim of this study is to investigate the prevalence of serious mental illness among pregnant women from 2016 to 2019. This was calculated by the proportion of pregnant women with serious

mental illness and those without from 2016 to 2019 in this national survey data. Descriptive statistics of all covariates, including the predictors and confounders, were also calculated to show if there are obvious differences in the demographic features and past month illicit drug use between those with and without serious mental illness.

The second aim of this study is to investigate how pregnant women with serious mental illness differ from those without serious mental illness in socioeconomic status, or if socioeconomic status was associated with serious mental illness in pregnant women, after adjusting for confounders. Before investigating this issue, it is important to first address the problem of confounding. As indicated above, the demographic features and past month illicit drug use were potential confounders, and no interaction term would be included due to a lack of literature support. In this study, these covariates would be adjusted using propensity score and inverse probability weighting. Since there were two predictors, income level and employment status, two propensity scores were calculated respectively for each individual by using logistic regression. The top 5% and bottom 5% of the estimated propensity scores were truncated to account for potential impact on the results.

Permutation tests for income level and employment status were first performed to assess their associations with the outcome and test the significance of these associations without considering the confounders. Next, two estimated differences in the appearance of serious mental disease between different income levels and different employment status were calculated, considering the inverse probability weighting using the propensity scores stated above. Re-randomization was then conducted, and the estimated differences were compared to the histograms obtained from re-randomization.

Two additional analyses were conducted as confirmatory analyses. The first one, while using the same inverse probability weighting and propensity score, used bootstrapping to obtain confidence intervals of the two estimated differences. The second analysis used a single logistic regression model, including all covariates, to estimate the coefficients of the two predictors, and used robust standard error to obtain confidence intervals. Wald tests were conducted on the two coefficients respectively to test if these coefficients were equal to 0.

Propensity scores, IPW analysis, bootstrapping, and logistic regressions were all conducted in R (3.6.1). 90% confidence intervals were reported and p-values less than 0.1 (2-sided) were considered statistically significant.

Results

Aim 1

Table 1 shows sample characteristics and SMI prevalence for the entire sample of pregnant women. SMI, with a score larger than 13, was estimated at 8.3% for the entire sample (n=229). Of those exhibiting SMI, approximately 79% came from families with an annual income less than $50,000, and 52% were unemployed in the past week during the survey. Most of them (38.9%) were in good health and only had some high school education (37.1%).

	K6<13	K6>=13	Overall
	(N=2525)	(N=229)	(N=2754)
Total Family Income			
< $50,000	1395 (55.2%)	181 (79.0%)	1576 (57.2%)
>= $50,000	1130 (44.8%)	48 (21.0%)	1178 (42.8%)
Past Week Working Status			
Unemployed	1019 (40.4%)	119 (52.0%)	1138 (41.3%)
Employed	1506 (59.6%)	110 (48.0%)	1616 (58.7%)
Overall Health			
Excellent	809 (32.0%)	25 (10.9%)	834 (30.3%)
Very good	971 (38.5%)	66 (28.8%)	1037 (37.7%)
Good	625 (24.8%)	89 (38.9%)	714 (25.9%)
Fair/Poor	120 (4.8%)	49 (21.4%)	169 (6.1%)
Education Categories			
< High school	322 (12.8%)	49 (21.4%)	371 (13.5%)
High school grad	668 (26.5%)	85 (37.1%)	753 (27.3%)
Some college/assoc	788 (31.2%)	80 (34.9%)	868 (31.5%)
College grad	747 (29.6%)	15 (6.6%)	762 (27.7%)
Race			
White	1370 (54.3%)	118 (51.5%)	1488 (54.0%)
Afr Am	387 (15.3%)	44 (19.2%)	431 (15.7%)
Native Am	43 (1.7%)	11 (4.8%)	54 (2.0%)
Native Hawaiian/Pacific Isl	13 (0.5%)	3 (1.3%)	16 (0.6%)
Asian	120 (4.8%)	1 (0.4%)	121 (4.4%)
1+ Race	83 (3.3%)	18 (7.9%)	101 (3.7%)
Hispanic	509 (20.2%)	34 (14.8%)	543 (19.7%)
Age Category			
18-25	1137 (45.0%)	168 (73.4%)	1305 (47.4%)

	K6<13	K6>=13	Overall
26-34	1094 (43.3%)	50 (21.8%)	1144 (41.5%)
35-49	294 (11.6%)	11 (4.8%)	305 (11.1%)
Marital Status			
Married	1414 (56.0%)	54 (23.6%)	1468 (53.3%)
Widowed	9 (0.4%)	0 (0%)	9 (0.3%)
Divorced/Separated	125 (5.0%)	14 (6.1%)	139 (5.0%)
Never married	977 (38.7%)	161 (70.3%)	1138 (41.3%)
Past Month Illicit Drug Use			
No	2362 (93.5%)	174 (76.0%)	2536 (92.1%)
Yes	163 (6.5%)	55 (24.0%)	218 (7.9%)

The sample consisted of 54% White non-Hispanic pregnant women, followed by 19.7% Hispanic, and 14.8% African American. Most women (47.4%) were between the ages of 18 and 25, while the rate of SMI for pregnant women in this age group was 73.4%, which is the highest among other age groups. The pregnant women who never got married took 41.3% of the sample, while the rate of SMI in this group was 70.3%. Those who were married took 53.3%, whereas the rate of SMI in this group was only 23.6%. As for illicit drug use in the past month, most women (92.1%) answered 'No', but the rate of 'Yes' among the SMI group (24%) is much higher than the rate of 'Yes' among the non-SMI group (6.5%)

Aim 2

We first checked to what extent the two income groups and the two past-week employment status groups differed in SMI prevalence using permutation tests without considering confounders. Figure 1 shows the permutation results of SMI prevalence difference between income and employment groups respectively, where the black line is the density of permuted difference, and the red vertical line is the observed difference. We estimated that the SMI prevalence is significantly higher ($p < 0.1$) for pregnant women in the low-income group (annual income < \$50,000) compared to the high-income group (annual income >= \$50,000). We also conclude that the SMI prevalence is significantly higher ($p < 0.1$) in the unemployment group compared to the employment group.

Considering the confounding effects, we used inverse probability weighting (IPW) to do the correction. As suggested, all potential confounding variables related to SMI were used for estimating the propensity score. Since income level and employment status are both binary variables, we applied logistic regression to estimate the propensity score. Figure 2 showed the density of propensity scores for income and employment respectively, and the corresponding density of weights for two variables are shown in Figure 3. Figure 4 showed the density (black line) of SMI difference between

two groups and the observed difference (red line). After the adjustment of race, education, marital status, health status, past-year illicit drug abuse, we don't have strong evidence to conclude that there is a significant difference ($p = 0.147$) in SMI prevalence between high- and low-income groups, nor ($p = 0.576$) between unemployment and employment groups.

We did further confirmatory analysis using bootstrap and logistic regression with SMI as the binary outcome, income and employment as the predictors, and all the other variables as confounders. The bootstrap result for income groups (90% CI: -0.043975, -0.00087) contradicted with the previous result, which indicates a significant difference in SMI prevalence. As for logistic regression, it shows there is a significant difference ($p < 0.1$) between different income groups as well.

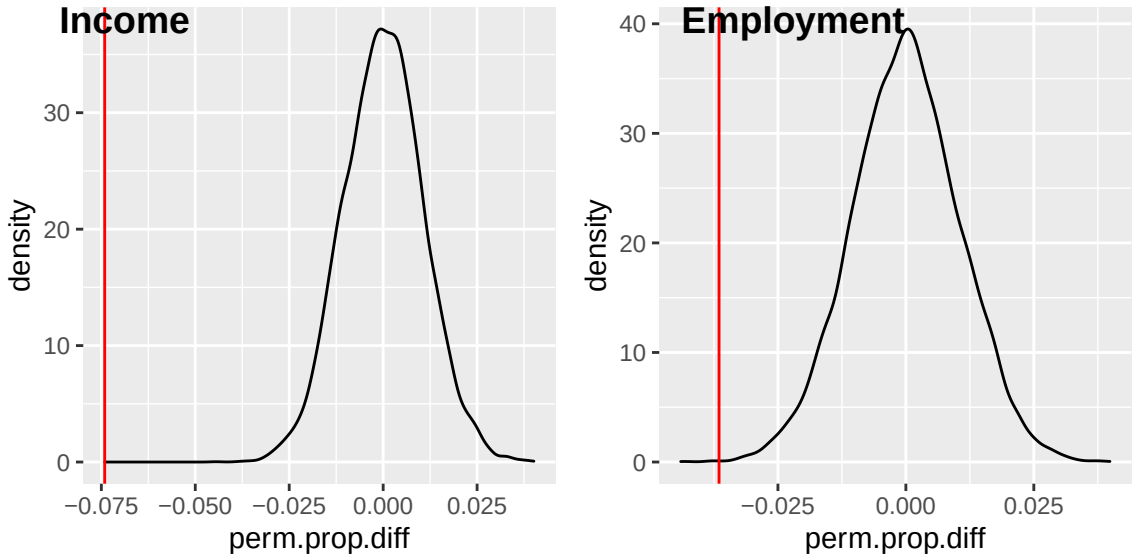

Figure 1: Permutated density of SMI difference and observed difference for income and employment

Figure 2: Propensity score density of SMI for income and employment

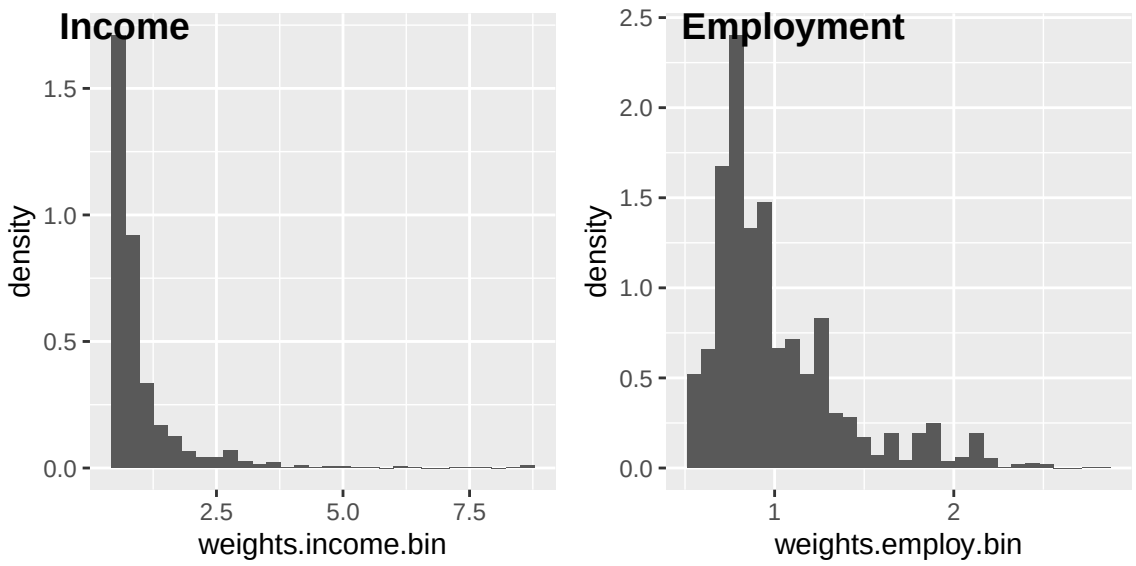

Figure 3: Weights density of SMI for income and employment

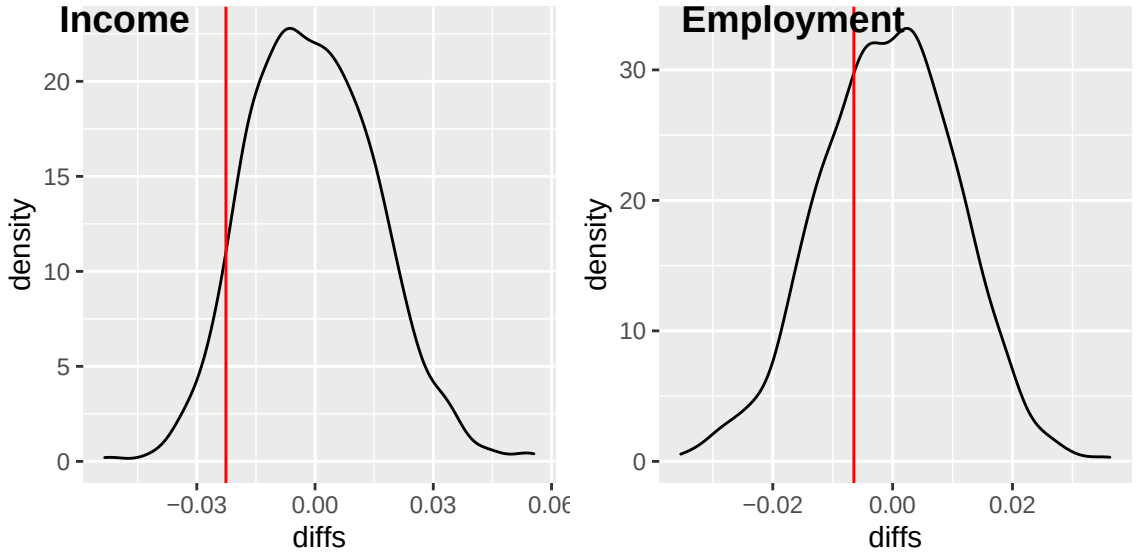

Figure 4: Permutated density of SMI difference after IPW and observed difference for income and employment

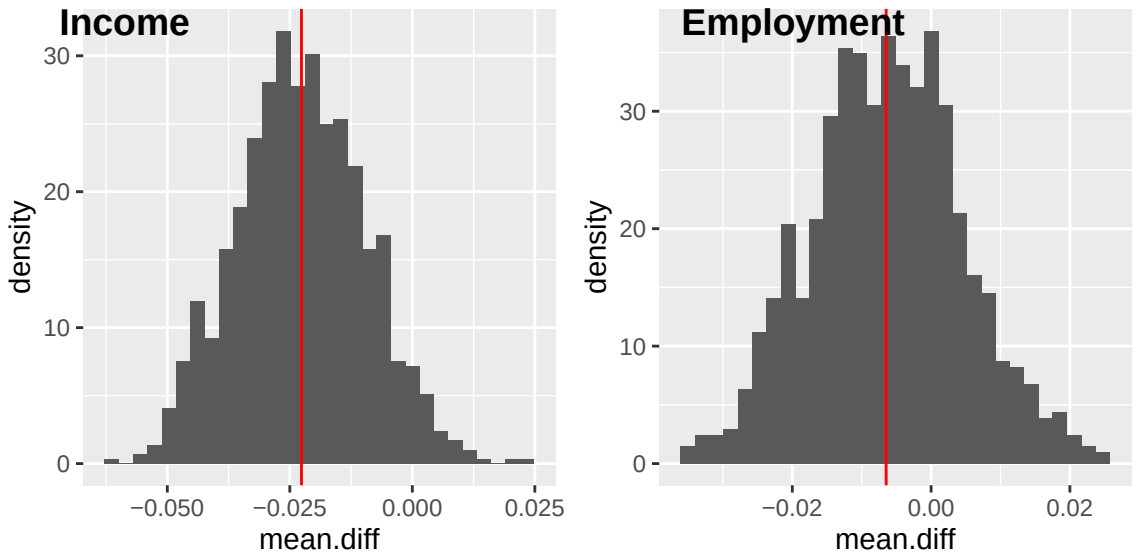

Figure 5: Bootstrapped SMI difference after IPW and observed difference for income and employment

Discussion

In the present study, we found out that the prevalence of serious mental illnesses in pregnant women from 2016 to 2019 to be 8.3%. Compared to the 5.6% of adults who were diagnosed with serious mental illness in 2020 (SAMHSA report, 2020), the percentage of pregnant women with serious mental illness is relatively at higher risk than the general population. This is a call to raise more public awareness, to advocate for better mental healthcare both in the health care delivering sector and also the strategic interventions to help pregnant women in accessing mental health care.

We are interested in whether there is an association between pregnant women with serious mental illness and income and work status. By looking at the mean difference ignoring the other confounders in a permutation test, we could observe that pregnant women with higher income and currently employed are less likely to have serious mental illnesses than those with lower income and unemployed. Based on our literature reviews, it is quite likely that income and employment are correlated with other features that are themselves correlated with serious mental illnesses. We used IPW with randomization to test the estimated differences in appearance of serious mental disease between different income levels with correction for the confounding effects, the result indicates the estimated differences between different income levels are not statistically significant ($p = 0.147$). However, in our confirmatory analysis using the bootstrap to capture the 90% confidence interval of the estimated difference between different income levels suggested the different result as the IPW method. It might be due to our choice of 90% confidence interval, with a significance level of 5% might give us a consistent result with the IPW re-randomization method. Further investigation is needed to have a full explanation on the contradicting results from the two methods. Our final confirmatory analysis with a logistic regression adjusting for all the confounding variables, sug-

gested that pregnant women in the family with a higher total income have a lower chance of having serious mental illness than in a lower income family. Therefore, it is essential to seek help for those with serious mental illness in pregnancy, especially when they are in a lower income family. The community needs to pay more attention to providing the necessary support and guidance for those who are seeking help.

Another limitation of our study is that we originally planned to run inverse probability weighting (IPW) on income and working status as multi-level variables. However, for the time being and the complicity in dealing with multi-level IPW and the potential problems with interpreting the results, we decided to dichotomize the two variables. By doing this, we realized that the dichotomization may let us lose some information on the association between the outcome and our covariates. Furthermore, due to the limited understanding of the survey data, we were not sure whether the missing values in the survey are missing at random or not, so we decided to treat it as missing completely at random for simplicity.

Despite these limitations above, this study enriches the current research on the association between serious mental illness and socioeconomic status in pregnant women. Future research should involve more variables that might improve the current study, such as performing IPW on income and working status as multi-level rather than binary variables or have a better way to deal with the missing data in the survey data. The data is from the National Survey on Drug Use and Health (NSDUH), so future research should also emphasize the drug use in pregnancy. These data are cross-sectional, so we were not able to track the changes of the mental health status over time. The survey data for 2020 was released recently, it would be a very interesting project to compare the prevalence before and after the COVID-19 pandemic.

References

Cook, C. A. L., Flick, L. H., Homan, S. M., Campbell, C., McSweeney, M., & Gallagher, M. E. (2010). Psychiatric disorders and treatment in low-income pregnant women. Journal of women's health, 19(7), 1251-1262.

Kitsantas, P., Aljoudi, S., Adams, A., & Booth, E. (2020). Prevalence and correlates of suicidal behaviors during pregnancy: Evidence from the National Survey on Drug Use and Health. Archives of Women's Mental Health, 24(3), 473-481.

Luciano, A., Nicholson, J., & Meara, E. (2014). The economic status of parents with serious mental illness in the United States. Psychiatric rehabilitation journal, 37(3), 242.

Salameh, T. N., Hall, L. A., Crawford, T. N., Staten, R. R., & Hall, M. T. (2020). Trends in mental health and substance use disorders and treatment receipt among pregnant and nonpregnant women in the United States, 2008–2014. Journal of Psychosomatic Obstetrics & Gynecology, 41(4), 298-307.

Prochaska, J. J., Sung, H. Y., Max, W., Shi, Y., & Ong, M. (2012). Validity study of the K6 scale as a measure of moderate mental distress based on mental health treatment need and utilization. International journal of methods in psychiatric research, 21(2), 88–97

Substance Abuse and Mental Health Services Administration. (2021). Key substance use and mental health indicators in the United States: Results from the 2020 National Survey on Drug Use and Health (HHS Publication No. PEP21-07-01-003, NSDUH Series H-56). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/

Appendix

```
library(tidyverse)
library(ggplot2)
library(ggpubr)
knitr::opts_chunk$set(echo = F)
# Data prep
# Load Data & select interest variables
load("./data/NSDUH_2016.RData")
load("NSDUH_2017.RData")
load("NSDUH_2018.RData")
load("NSDUH_2019.RData")

preg2016 <- PUF2016_022818 %>% filter(pregnant == 1)
preg2016 <- preg2016 %>% select(spdmon,NEWRACE2,CATAG3,irmarit,HEALTH2,eduhighcat,WRKSTATWK2,income,illmon)
preg2016 <- na.omit(preg2016)

preg2017 <- PUF2017_100918 %>% filter(pregnant == 1)
preg2017 <- preg2017 %>% select(spdmon,NEWRACE2,CATAG3,irmarit,HEALTH2,eduhighcat,WRKSTATWK2,income,illmon)
preg2017 <- na.omit(preg2017)

preg2018 <- PUF2018_100819 %>% filter(pregnant == 1)
preg2018 <- preg2018 %>% select(spdmon,NEWRACE2,CATAG3,irmarit,HEALTH2,eduhighcat,WRKSTATWK2,income,illmon)
preg2018 <- na.omit(preg2018)

preg2019 <- PUF2019_100920 %>% filter(pregnant == 1)
preg2019 <- preg2019 %>% select(spdmon,NEWRACE2,CATAG3,irmarit,HEALTH2,eduhighcat,WRKSTATWK2,income,illmon)
preg2019 <- na.omit(preg2019)

# Combine data
preg2016 <- read.csv("./Data/preg2016.csv")
preg2017 <- read.csv("./Data/preg2017.csv")
preg2018 <- read.csv("./Data/preg2018.csv")
preg2019 <- read.csv("./Data/preg2019.csv")

preg.dat <- rbind(preg2016, preg2017, preg2018, preg2019)
## Imputation of interest variable
preg.dat <- preg.dat %>%
  mutate(
    jobstat = case_when(
      WRKSTATWK2 <= 3 ~ 1,
      WRKSTATWK2 == 98 ~ 5,
      WRKSTATWK2 == 5 ~ 3,
      WRKSTATWK2 == 7 ~ 4,
      TRUE ~ 2
    )
  )
preg.dat$WRKSTATWK2 <- NULL
preg.dat$income01 <- ifelse(preg.dat$income == 2 | preg.dat$income == 1, 0,1 )
preg.dat$employ01 <- ifelse(preg.dat$jobstat == 5, NA, preg.dat$jobstat)
preg.dat <- na.omit(preg.dat)
preg.dat$employ01 <- ifelse(preg.dat$jobstat ==1, 1, 0)
preg.dat$income <- NULL
preg.dat$jobstat <- NULL

# Save data for further use
write.csv(preg.dat, file = "preg_all.csv")
## Descriptive Statistics
preg_all <- read.csv("preg_all.csv")[,-1]
library(table1)
preg.dat.table <- preg_all

preg.dat.table$spdmon <- factor(preg.dat.table$spdmon, levels = c(0,1),
                                labels = c("K6<13","K6>=13"))
label(preg.dat.table$spdmon) <- "Past Month K6"

preg.dat.table$NEWRACE2 <- factor(preg.dat.table$NEWRACE2, levels = c(1:7),
                                labels = c("White","Afr Am", "Native Am","Native Hawaiian/Pacific Isl","Asian","1+ Race","Hispanic"))
label(preg.dat.table$NEWRACE2) <- "Race"

preg.dat.table$CATAG3 <- factor(preg.dat.table$CATAG3, levels = c(2,3,4),
                              labels = c("18-25","26-34","35-49"))
label(preg.dat.table$CATAG3) <- "Age Category"

preg.dat.table$irmarit <- factor(preg.dat.table$irmarit, levels = c(1:4),
                              labels = c("Married","Widowed","Divorced/Separated","Never married"))
label(preg.dat.table$irmarit) <- "Marital Status"

preg.dat.table$HEALTH2 <- factor(preg.dat.table$HEALTH2, levels = c(1:4),
                              labels = c("Excellent","Very good","Good","Fair/Poor"))
label(preg.dat.table$HEALTH2) <- "Overall Health"

preg.dat.table$eduhighcat <- factor(preg.dat.table$eduhighcat, levels = c(1:4),
                                labels = c("< High school","High school grad","Some college/assoc","College grad"))
label(preg.dat.table$eduhighcat) <- "Education Categories"

preg.dat.table$employ01 <- factor(preg.dat.table$employ01, levels = c(0,1),
                                labels = c("Unemployed","Employed"))
label(preg.dat.table$employ01) <- "Past Week Working Status"

preg.dat.table$income01 <- factor(preg.dat.table$income01, levels = c(0,1),
                                  labels = c("< $50,000",">= $50,000"))
label(preg.dat.table$income01) <- "Total Family Income"
```

```r
preg.dat.table$illmon <- factor(preg.dat.table$illmon, levels = c(0,1),
                                labels = c("No","Yes"))
label(preg.dat.table$illmon) <- "Past Month Illicit Drug Use"

table1(~income01+employ01+HEALTH2+eduhighcat+NEWRACE2+CATAG3+irmarit+illmon|spdmon, data = preg.dat.table, caption = 'Sample characteristics of pregnant women')

# prevalence: 229/2754=0.832
prop.diff.income = mean(preg_all$spdmon[preg_all$income01 == 1]) -  mean(preg_all$spdmon[preg_all$income01 == 0])
prop.diff.employ = mean(preg_all$spdmon[preg_all$employ01 == 1]) - mean(preg_all$spdmon[preg_all$employ01 == 0])

do.one <- function(outcome, label){
  perm.label <- sample(label)
  return(mean(outcome[perm.label == 1]) - mean(outcome[perm.label == 0]))
}

set.seed(1)
sampling.dist.income = with(preg_all, replicate(1e4,do.one(preg_all$spdmon,preg_all$income01)))
sampling.dist.employ = with(preg_all, replicate(1e4,do.one(preg_all$spdmon,preg_all$employ01)))

perm_income <- ggplot(data.frame(perm.prop.diff = sampling.dist.income), aes(x = perm.prop.diff, y=..density..)) +
  geom_density() +
  geom_vline(xintercept = prop.diff.income, color = "red")

perm_employ <- ggplot(data.frame(perm.prop.diff = sampling.dist.employ), aes(x = perm.prop.diff, y=..density..)) +
  geom_density() +
  geom_vline(xintercept = prop.diff.employ, color = "red")

ggarrange(perm_income, perm_employ, labels = c('Income', 'Employment'), nrow = 1, ncol = 2)

# Permutation p-value
# mean(abs(sampling.dist.income)>abs(prop.diff.income))
# mean(abs(sampling.dist.employ)>abs(prop.diff.employ))
propen.model.income.bina = glm(income01~illmon+eduhighcat+HEALTH2+CATAG3+irmarit+NEWRACE2, data = preg_all,
                               family = binomial)
propen.model.employ.bina = glm(employ01~illmon+eduhighcat+HEALTH2+CATAG3+irmarit+NEWRACE2, data = preg_all,
                               family = binomial)

propensities.income.bina = predict(propen.model.income.bina, data = preg_all, type = "response")
propensities.employ.bina = predict(propen.model.employ.bina, data = preg_all, type = "response")

pro_income <- ggplot(data.frame(propensities=propensities.income.bina, income = as.factor(preg_all$income01)),
       aes(x = propensities, y = ..density.., color = income)) + geom_density()
pro_employ <- ggplot(data.frame(propensities=propensities.employ.bina, employ = as.factor(preg_all$employ01)),
       aes(x = propensities, y = ..density.., color = employ)) + geom_density()

ggarrange(pro_income, pro_employ, labels = c('Income', 'Employment'), nrow = 1, ncol = 2)

trunc.propen.income.bina = propensities.income.bina %>% pmin(0.95) %>% pmax(0.05)
trunc.propen.employ.bina = propensities.employ.bina %>% pmin(0.95) %>% pmax(0.05)
npat = nrow(preg_all)

weights.income.bin = rep(0, npat)
weights.employ.bin = rep(0, npat)

representative.propen.income = sum(preg_all$income01)/npat
representative.propen.employ = sum(preg_all$employ01)/npat
actual.propen.income.bina = trunc.propen.income.bina
actual.propen.employ.bina = trunc.propen.employ.bina

income.ind = which(preg_all$income01 == 1)
employ.ind = which(preg_all$employ01 == 1)

weights.income.bin[income.ind] = representative.propen.income/actual.propen.income.bina[income.ind]
weights.income.bin[-income.ind] = (1-representative.propen.income)/(1-actual.propen.income.bina[-income.ind])
weights.employ.bin[employ.ind] = representative.propen.employ/actual.propen.employ.bina[employ.ind]
weights.employ.bin[-employ.ind] = (1-representative.propen.employ)/(1-actual.propen.employ.bina[-employ.ind])

w_income <- ggplot(data.frame(weights = weights.income.bin), aes(x=weights.income.bin, y =..density..)) +
  geom_histogram()
w_employ <- ggplot(data.frame(weights = weights.employ.bin), aes(x=weights.employ.bin, y =..density..)) +
  geom_histogram()

ggarrange(w_income, w_employ, labels = c('Income', 'Employment'), nrow = 1, ncol = 2)

income.prop.est = with(preg_all, mean((weights.income.bin*spdmon)[income.ind]))
noincome.prop.est = with(preg_all,mean((weights.income.bin*spdmon)[-income.ind]))
diff.income.est = income.prop.est - noincome.prop.est

employ.prop.est = with(preg_all,mean((weights.employ.bin*spdmon)[employ.ind]))
noemploy.prop.est = with(preg_all,mean((weights.employ.bin*spdmon)[-employ.ind]))
diff.employ.est = employ.prop.est - noemploy.prop.est
do.one.propen.binary <- function(outcome, propen){
  n <- length(outcome)
  label <- rbinom(n,1,propen)

  weights <- rep(0,n)
  representative <- mean(label)
  actual <- propen
  ind.t <- which(label == 1)
  weights[ind.t] <- (representative/actual)[ind.t]
  weights[-ind.t] <- ((1-representative)/(1-actual))[-ind.t]

  return(mean((weights*outcome)[ind.t]) - mean((weights*outcome)[-ind.t]))
}

set.seed(1)
```

```
rerandomized.diffs.income = replicate(1e3, do.one.propen.binary(preg_all$spdmon, trunc.propen.income.bina))
rerandomized.diffs.employ = replicate(1e3, do.one.propen.binary(preg_all$spdmon, trunc.propen.employ.bina))

ipw_income <- ggplot(data.frame(diffs = rerandomized.diffs.income), aes(x = diffs, y = ..density..)) +
  geom_density() +
  geom_vline(xintercept = diff.income.est, color = "red")
ipw_employ <- ggplot(data.frame(diffs = rerandomized.diffs.employ), aes(x = diffs, y = ..density..)) +
  geom_density() +
  geom_vline(xintercept = diff.employ.est, color = "red")

ggarrange(ipw_income, ipw_employ, labels = c('Income', 'Employment'), nrow = 1, ncol = 2)

# mean(abs(rerandomized.diffs.income)>abs(diff.income.est))
# mean(abs(rerandomized.diffs.employ)>abs(diff.employ.est))
calc_weighted_outcome <- function(outcome, label, props){
  weights <- rep(0, length(outcome))

  representative.propen <- mean(label)
  actual.propen <- props

  treat.ind <- which(label == 1)
  weights[treat.ind] <- representative.propen/actual.propen[treat.ind]
  weights[-treat.ind]<- (1 - representative.propen)/(1- actual.propen[-treat.ind])

  weighted.outcome <- weights*outcome

  return(weighted.outcome)
}

calc_stat_weighted <- function(weighted.outcome, label){
  return(mean(weighted.outcome[label == 1]) - mean(weighted.outcome[label == 0]))
}

do_one_income <- function(dat){
  resample.inds <- sample(1:nrow(dat), replace=TRUE)
  resample.dat <- dat[resample.inds,]

  propen.model <- glm(income01~illmon+eduhighcat+HEALTH2+CATAG3+irmarit+NEWRACE2, data = resample.dat,
                      family = binomial)
  propensities <- predict(propen.model, data = resample.dat, type = "response")
  trunc.prop <- propensities %>% pmax(0.05) %>% pmin(0.95)

  weighted.outcome.resamp <- calc_weighted_outcome(resample.dat$spdmon,
                                                   resample.dat$income01,
                                                   trunc.prop)
  mean.diff <- calc_stat_weighted(weighted.outcome.resamp, resample.dat$income01)
  return(mean.diff)
}

do_one_employ <- function(dat){
  resample.inds <- sample(1:nrow(dat), replace=TRUE)
  resample.dat <- dat[resample.inds,]

  propen.model <- glm(employ01~illmon+eduhighcat+HEALTH2+CATAG3+irmarit+NEWRACE2, data = resample.dat,
                      family = binomial)
  propensities <- predict(propen.model, data = resample.dat, type = "response")
  trunc.prop <- propensities %>% pmax(0.05) %>% pmin(0.95)

  weighted.outcome.resamp <- calc_weighted_outcome(resample.dat$spdmon,
                                                   resample.dat$employ01,
                                                   trunc.prop)
  mean.diff <- calc_stat_weighted(weighted.outcome.resamp, resample.dat$employ01)
  return(mean.diff)
}


mean.diff.est.income <- calc_stat_weighted(calc_weighted_outcome(preg_all$spdmon, preg_all$income01, trunc.propen.income.bina),
                                           preg_all$income01)
mean.diff.est.employ <- calc_stat_weighted(calc_weighted_outcome(preg_all$spdmon, preg_all$employ01, trunc.propen.employ.bina),
                                           preg_all$employ01)

set.seed(1)
boot.dist.income <- replicate(1e3, do_one_income(preg_all))
boot.dist.employ <- replicate(1e3, do_one_employ(preg_all))

boot_income <- ggplot(data.frame(mean.diff = boot.dist.income), aes(x = mean.diff, y=..density..)) + geom_histogram() +
  geom_vline(xintercept=mean.diff.est.income, color="red")
boot_employ <- ggplot(data.frame(mean.diff = boot.dist.employ), aes(x = mean.diff, y=..density..)) + geom_histogram() +
  geom_vline(xintercept=mean.diff.est.employ, color="red")

ggarrange(boot_income, boot_employ, labels = c('Income', 'Employment'), nrow = 1, ncol = 2)

distance.U.L.income <- quantile(boot.dist.income, c(0.05,0.95)) - mean.diff.est.income
distance.U.L.employ <- quantile(boot.dist.employ, c(0.05,0.95)) - mean.diff.est.employ

# (CI.income <- mean.diff.est.income - distance.U.L.income[2:1])
# (CI.employ <- mean.diff.est.employ - distance.U.L.employ[2:1])
log.model <- glm(spdmon~income01+employ01+HEALTH2+eduhighcat+NEWRACE2+CATAG3+irmarit+illmon, data = preg_all, family = binomial)
# summary(log.model)

library("sandwich")
coef <- log.model$coef
rob_se <- sqrt(diag(vcovHC(log.model, type = "HC0")))
# (ci_income <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2])
# (ci_employ <- coef[3] + c(0, qnorm(c(0.025, 0.975))) * rob_se[3])
```