---
title: "The Effects Of Smoking & Mother's Age on Pregnancy & Newborn's Health"
author: "Alexandria Lange"
description: "Gov 50 Final Project"

output:
  distill::distill_article:
    self_contained: false

Milestone 1: Project thoughts

For my final project, I will be studying the effects of certain habits that pregnant mothers participate in that may affect their overall health and how that impacts the child. I found a great data set that includes birth statistics from 1000 randomly selected births in 2014. These statistics are mother's age, father's age, mother's race, number of weeks pregnant, amount of weight gained during pregnancy, weight of baby, gender of baby, sex of baby, the maturity of the mother (younger or older), if the baby was a preemie or was carried to full term, if the birth weight was low, smoking habits (whether or not they smoke), and if the mother was white or not. Although there are a lot of interesting questions that can be explored in this data set, I hope to focus on some of the harmful effects of smoking on a baby's health and on the pregnancy in general. I would also like to compare the effects of habits and the effects of the mother's age, as my mother was very aware of the increased risks to my health as she got pregnant with me later than most women. I'm interested in comparing these two as they are both known to be significant factors during pregnancy.

Milestone 2: Proposal

After analyzing the data set and considering possible research questions, I have decided to investigate the effects of smoking habits on the probabilities of a pregnant female being able to carry to term and of her baby being born with a low birth weight (less than 5lbs 8 ounces). I suspect there will be a higher rate of babies born prematurely and a higher rate of babies born with a low birth weight in the group of mothers who had a habit of smoking compared to the group of mothers who did not.

While smoking is harmful to an individual's health in any circumstance, its harmful effects go beyond the individual during pregnancy and have the ability to impact a developing fetus' health. The Centers for Disease Control and Prevention (CDC) has published that smoking during pregnancy can cause tissue damage in the lung(s) and the brain of the unborn baby, as well as increase the odds of the child having a cleft lip. There is also mention that smoking during pregnancy leads to preterm labor and low birth weight. In the following study,I will test both of these statements to see whether they hold true.

Therefore, my research conjecture is that smoking makes premature birth and/or low fetal birth weight more likely, in line with the claims of the CDC. The data set I will use to perform my analysis consists of a random sample of 1000 birth cases that occurred in the US in 2014. Given that random sampling was used, my results should be generalizable to all pregnant women in the US.

The null hypothesis (H_{0}) is that smoking does not make premature birth and/or low fetal birth weight more likely. The alternative hypothesis (H_{a}) is that there is positive association between smoking habits and the likelihood of premature birth and/or low fetal birth weight.

In order to test my hypotheses, I will first create binary variables out of the `preemie` variable (`ifpreemie`) and out of the `lowbirthweight` variable (`iflow`), assigning "1" to if they were a preemie and if they were born with a low birth weight. From there, I will be able to calculate the average proportion of preemies among smokers and non-smokers, as well as the average proportion of babies born with a low birth weight within each group. If the average proportion of preemies and babies born with a low birth weight are each higher within the group of mothers with smoking habits than the group of mothers that did not have smoking habits (during pregnancy), then this will help support the alternative hypothesis.

I will also run a regression to see if there is an association between smoking and preemies as well as smoking and being born with a low birth weight. It is important to keep in mind, however, that while my data may suggest a conclusion, I will not be able to establish causality because there are many confounding factors that could affect the chances of preterm labor and whether or not a baby is born with a low birth weight.

As for comparing the effects of smoking habits and those of the mother's age, I will do so by doing multiple variable regressions. The independent variables in these regressions will be habits (`habit`) and mother's age (`fage`) and the dependent variable in the first one will be the binary variable I created for if the baby was a preemie (`ifpreemie`) and in the second will be the binary variable I created for if the baby had a low birth weight (`iflow`). The linear regression coefficient that results from each will allow me to guage differences in impact of each independent variable on the respective dependent variable.

The null hypothesis for this section (my second null hypothesis) is that an increase in the mother's age does not make premature birth and/or low fetal birth weight more likely. The alternative hypothesis to this is that there is a positive connection between an increase in mother's age and the likelihood of premature birth and/or low fetal birth weight.

I also chose to include a box plot to show the distribution in baby's weights (in pounds) split up by smoking habits of the mother to show a more general visualization of the effects of smoking on a baby's weight (not just whether or not the baby had a low birth weight/ was under 5lbs 8ounces).

The research design of this project is a "cross sectional design", as I will be comparing outcomes of treated and control groups at one point in time. The treatment in this case is smoking, so the treatment group includes the mothers who have smoking habits and the control group is the group of mothers who do not have smoking habits. The outcomes are whether or not their baby was born prematurely and if they were born with a low birth weight.

Clarification:
Null Hypothesis #1:smoking does not make premature birth and/or low fetal birth weight more likely

Null Hypothesis #2:an increase in the mother's age does not make premature birth and/or low fetal birth weight more likely

Milestone 3:Data Visualizations

```{r}
library(tidyverse)
library(broom)
library(ggplot2)
library(readr)
library(infer)
births14 <- read_csv("~/Downloads/births14 (1).csv")

births14<-rename(births14, preemie= premie)
```

```{r}
births14<-births14|>
  mutate(ifpreemie=ifelse(preemie == "premie", 1, 0))

habit_preemie_mean<-births14|>
  group_by(habit)|>
  remove_missing()|>
summarise(preemie_mean= mean(ifpreemie))

knitr::kable(digits = 2, habit_preemie_mean)
```

```{r}
habit_preemie<-births14|>
  group_by(habit)|>
  remove_missing()|>
summarise(preemie_mean= mean(ifpreemie))|>
```

```
  ggplot(mapping = aes(x= habit,
                       y=preemie_mean))+
  geom_col(fill=c("lightskyblue","purple1"))+
  labs(title="Average Proportion of Preemies by US Mothers' Smoking Habits (2014)",
         x= "Smoking Habits",
       y= "Proportion of Preemies") +
  theme(plot.title=element_text(hjust=0.5))

habit_preemie

```

Interpretation of Table & Graph

In the group of babies born to mothers who had smoking habits, roughly **0.24 (24%)** of them were born prematurely, while in the group of babies born to mothers who did not have smoking habits, roughly **0.11 (11%)** were born prematurely.

The proportion of babies that were born prematurely more than doubled when the mother smoked, compared to when the mother didn't. This (preliminarily) supports the first part of my first alternative hypothesis, which is that smoking during pregnancy makes premature birth more likely, as per the CDC.

Regression of Premature Births Given Smoking Habits

```{r}
ifpreemie_habit<-lm(ifpreemie~habit, data= births14)
ifpreemie_habit
```

Interpretation of Linear Regression

0.1107 = 11.07% the average probability of a baby being born prematurely when the mother does not have a habit of smoking

0.1086 = 10.86% the average change (increase) in the probability of a baby being born prematurely when the habit of smoking is present in the mother

This shows that there is an increase in the probability of babies being born prematurely with the habit of smoking being present in the mother, thus supporting the first alternative hypothesis.

```{r}
summary_ifpreemie_habit<-modelsummary::modelsummary(ifpreemie_habit,
                         statistic = c("s.e.= {std.error}",
                                       "p={p.value}"),
                          gof_map= c("ifpreemie", "habit"))
summary_ifpreemie_habit
```

Interpretation of the P-Value for Preemie

The P value is **less than 0.001**. This is highly statistically significant and tells us that observing a regression coefficient of 0.109 or more extreme is very unlikely under the first null hypothesis.Therefore, there is evidence to support the first part of the first alternative hypothesis that there is positive association between smoking habits and the likelihood of premature birth, and we can reject the null hypothesis.

Regression of Preemies Given Smoking Habits & Mother's Age

```{r}
```

```
preemie_multi<-lm(ifpreemie~habit+ fage, data= births14)
preemie_multi
```

Interpretation of Linear Regression

0.103766= about 10.38%=the average change (increase) in the probability of a baby being born prematurely when the habit of smoking is present in the mother. This implies that the probability of a baby being born prematurely increases by about 10% when the mother has a habit of smoking (when the mothers are the same age).

0.003042= about 0.30%= the average change (increase) in the probability of a baby being born prematurely when the age of the mother is increased by one unit. This implies that the probability of a baby being born prematurely increases by about 0.3% when a mother's age increases by one year & the mother does not have a habit of smoking.

A greater average increase in the probability of a born prematurely in this data set with the presence of smoking in the mother than with a one unit increase in the mother's age indicates that smoking has a greater effect on the odds of a baby being born prematurely than a mother's age increase(by one year) does.

As for comparing the difference in habitsmoker from this multi regression where it is **0.103766** and the regression above where it is **0.1086**, they are fairly similar. The average change (increase) in the probability of a baby being born prematurely is only **0.004834 (0.4%)** smaller when the confounding factor of the mother's age is included. Thus, it shows that out of the two independent variables (smoking and mother's age), both increase the probability of a baby being born prematurely, but smoking habits has a much greater effect than an increase of mother's age by one year. The addition of the confounding factor of mother's age doesn't affect the regression value enough to dispute the significance of the impact of the mother's smoking habits on the likelihood of premature birth.

```{r}
summary_ifpreemie_multi<-modelsummary::modelsummary(preemie_multi,
                         statistic = c("s.e.= {std.error}",
                                       "p={p.value}"),
                          gof_map= c("ifpreemie", "habit", "fage"))
summary_ifpreemie_multi
```

Reporting & Interpreting P-values from Multi-Regression Model Summary
Smoking Habits of the Mother (habitsmoker) p-value= **0.004**. Since this value is less than 0.01, this value is highly significant and tells us that observing a regression coefficient of 0.104 or greater is very unlikely under the null hypothesis. Therefore, we would reject the 1st part of my 1st null hypothesis (that smoking does not make premature birth more likely).

Mother's Age (fage) p-value=**0.055**. Since this value is greater than 0.05 (but not by much), it is not very statistically significant (would be categorized as "statistically significant" if it was less than 0.05). Therefore, it is not enough to reject the first part of my second null hypothesis (that an increase in mother's age does not make premature birth more likely) and we would fail to reject the the first part of my second null hypothesis.

***

Effects of Habit on Baby Weight

```{r}
births14<-births14|>
  mutate(ifsmoker=ifelse(habit == "smoker", 1, 0))

habit_birthweight<-births14|>
  group_by(weight, habit)|>
  remove_missing()|>
```

```
  summarise(baby_weight= mean(weight))|>
  ggplot(mapping = aes(x= habit,
                       y=baby_weight))+
 geom_boxplot(fill = c("lightskyblue","purple1"))+
  labs(title = "Baby Weights (in pounds) Separated by Habits (Nonsmoker v. Smoker)",
 x= "Smoker Habit",
y= "Baby Weight (in pounds)")+
  theme(plot.title=element_text(hjust=0.5))
habit_birthweight
```

Interpretation of Boxplot

This boxplot shows the distribution of baby weights by habits of the mothers (smokers v non smokers). The median weight in the non smoker group is about 7.4 pounds, whereas the median weight for the smoker group is lower (about 6.8 pounds). As for the middle 50% of weights in both data sets, there is an overlap from about 6.25 pounds to 7.6 pounds. However, this overlap is about the 37th-75th percentile in weights of babies whose mothers did smoke overlapping with the 25th to (about) the 52nd percentile in weights of babies whose mothers did smoke. Therefore, the middle 50% of weights of babies born to smokers includes lower weights than the middle 50% of weights of babies born to non smokers. The range of the middle 50% (25th-75% percentile) of weights of babies born to smokers was about 5.8 to 7.6 lbs, while the range of the middle 50% of weights of babies born to nonsmokers ranged from about 6.25 to about 8 lbs. Further, the weights of the babies born to nonsmokers was more concentrated than those of babies born to smokers. As for the top 25th percentile in weight, the groups of moms who were non smokers included heavier baby weights than the group of moms who did smoke, with the 100th percentile in baby weight being about 10.2 pounds in the group of non smoking moms and about 8.9 lbs in the group of moms who did smoke.

As for the outliers, the lightest baby belonged to one whose mothers had smoking habits.

Overall, this graph indicates the increased likelihood of a child born to a mother who smokes having a lower birth weight than a child being born to a mother that does not smoke.

***

Effect of Habits on Low Birth Weight

```{r}
births14<-births14|>
  mutate(iflow=ifelse(lowbirthweight == "low", 1, 0))

habitlowbirthw_mean<-births14|>
  group_by(habit)|>
  remove_missing()|>
summarise(low_mean= mean(iflow))

knitr::kable(digits = 2, habitlowbirthw_mean)
```

```{r}
habit_lowbirthweight<-births14|>
  group_by(habit)|>
  remove_missing()|>
summarise(low_mean= mean(iflow))|>
  ggplot(mapping = aes(x= habit,
                       y=low_mean))+
  geom_col(fill=c("lightskyblue","purple1"))+
  labs(title="Average Proportion of Babies Born at a Low Birth Weight by
       US Mothers' Smoking Habits (2014)",
         x= "Smoking Habits",
       y= "Proportion of Babies Born at Low Birth Weight")+
  theme(plot.title=element_text(hjust=0.5))
habit_lowbirthweight

```

Interpretation of Chart and Graph:
0.6 (6%) of the babies that were born to mothers who were nonsmokers were classified as having a "low birth weight", whereas **0.23 (23%)** of the babies that were born to mothers that were smokers were classified as having a "low birth weight". The proportion of babies that were classified as having a "low birth weight" almost quadrupled when the mother smoked, in comparison to when the mother didn't. This supports the claim of the CDC that smoking during pregnancy increases the chances that the child will be born with a low birth weight.

Regression of Low Birth Weight Baby Given Smoking Habits

```{r}
iflow_habit<-lm(iflow~habit, data= births14)
iflow_habit
```

Interpretation of Linear Regression:

0.06228= about 6.23%=the average probability of a baby being born with a classified "low birth weight" when the mother does not have a habit of smoking

0.14824= about 14.82%=the average change (increase) in the probability of a baby being born with a classified "low birth weight" when the mother has a habit of smoking

There is an increase in the average probability of a baby being born with a classified "low birth weight" when the habit of smoking was present in the mother, thus supporting the first part of the second alternative hypothesis.

```{r}
summary_iflow_habit<-modelsummary::modelsummary(iflow_habit,
                        statistic = c("s.e.= {std.error}",
                                      "p={p.value}"),
                         gof_map= c("iflow", "habit"))
summary_iflow_habit
```

Interpretation of the P-Value for Low Birthweight

The P value for smoking habits of the mother (habitsmoker) is **less than 0.001**. This is highly statistically significant and tells us that observing a regression coefficient of 0.148 or more extreme is unlikely under the first part of the second null hypothesis, so we would reject it.Therefore, there is evidence to support the first part of the second alternative hypothesis that there is positive association between smoking habits and the likelihood of low birth weight.

Regression of Babies born at a Low Birth Weight Given Smoking Habits & Mother's Age

```{r}
iflow_multi<-lm(iflow~habit+fage, data= births14)
iflow_multi
```

Interpretation of Linear Regression:

0.124498= about 12.46%= the average change (increase) in the average probability of a baby being born with a classified "low birth weight" when the mother has a habit of smoking (when mothers are the same age)

This implies that the average probability of a baby being born prematurely increases by about 12% with the habit of smoking being present in the mother (when the mothers are the same age).

0.001422= about 0.14%= the average change (increase) in the probability of a baby being born with a classified "low birth weight" when the age of the mother is increased by one unit and the mother does not have a habit of smoking. This implies that the chances of a baby being born prematurely increases by about 0.1% when a mother's age increases by one year.

Both the average change in the probability of a baby being born with a classified "low baby weight" when age of the mother is increased by one unit and the habit of smoking is present in the mother are positive (holding the other independent variable constant). However, the average change in the probability is much greater when when the habit of smoking is present in the mother (of mothers the same age without smoking habits) than when the age of the mother being increased by one year is the only changing variable, signifying that smoking has a greater effect on whether or not a baby is born with a low birth weight than the age of the mother.

When comparing the linear regression coefficient of habitsmoker in this mutli regression when is it is **0.124498** and the regression above when it is **0.14824**, the average increase in probability of a baby being born at a low birth weight is only **0.023782 (about 2.38%)** smaller when the confounding factor of mother's age is not taken into account. Thus, it works to emphasize how much big of an impact smoking has on the probability of a child being born with a low birth weight even with this possible confounding factor being included.

```{r}
summary_iflow_multi<-modelsummary::modelsummary(iflow_multi,
                         statistic = c("s.e.= {std.error}",
                                       "p={p.value}"),
                          gof_map= c("iflow", "habit", "fage"))
summary_iflow_multi
```

Reporting & Interpretation of P-values
P value for Smoking Habits of the Mother (habitsmoker) is **less than 0.001**. Therefore, it is highly significant and tells us that observing a regression coefficient of 0.124 or greater is highly unlikely under the 2nd part of my 1st null hypothesis. Therefore, we would reject the 2nd part of my first my null hypothesis (that smoking does not make low birth weight more likely).

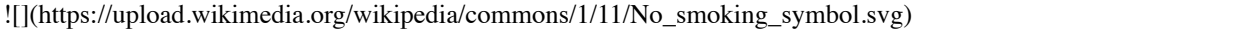
P value for Mother's Age (fage) is **0.250**. Since this value is greater than 0.1, it is not statistically significant. It means that a regression coefficient of 0.001 or greater would only happen in 25% of repeated samples if the null was true. Therefore, it occurs often enough that we would fail to reject the second part of my second null hypothesis (that an increase in the mother's age does make low birth weight more likely).

***

Conclusion

Overall, the (first) null hypothesis that smoking does not make premature birth and low birth weight more likely was rejected every time it was evaluated; the (second) null hypothesis that (an increase in) mother's age does not cause premature birth and low birth weight to be more likely failed to be rejected both times it was evaluated.

Thus, if a mother smoked, it was found to increase the probability of her baby being born prematurely and/or with a low birth weight. While smoking habits' effects on these probabilities remained somewhat constant with the inclusion of one possible confounding variable (mother's age/ "fage"), there are many other possible confounding factors that, if tested alongside smoking habits, may cause the effects of smoking on the two dependent variables mentioned previously to dwindle. Therefore, I am unable to confidently conclude causality between smoking habits and preterm labor and/or low birth weight.



Data from: [link](https://www.openintro.org/data/index.php?data=births14)

CDC claims from: [link](https://www.cdc.gov/tobacco/campaign/tips/diseases/pregnancy.html#:~:text=Smoking%20slows%20your%20baby's%20growth,babies%20often%20have%20health%20problems.&text=S