IDS 702
Modeling and Representation of Data

Maternal Smoking and Pre-term Birth

These days, it is widely understood that mothers who smoke during pregnancy risk exposing their babies to many health problems. This was not common knowledge fifty years ago. One of the first studies that addressed the issue of pregnancy and smoking was the Child Health and Development Studies, a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, CA. The original reference for the study is Yerushalmy (1964, *American Journal of Obstetrics and Gynecology,* pp. 505-518). The data and a summary of the study are in Nolan and Speed (2000, *Stat Labs,* Chapter 10) and can be found at the book's website. The data are in the Course Datasets folder on the IDS 702 Sakai course page.

There were about 15,000 families in the study. We will analyze a subset of the data, in particular 1236 male single births where the baby lived at least 28 days. The researchers interviewed mothers early in their pregnancy to collect information on socioeconomic and demographic characteristics, including an indicator of whether the mother smoked during pregnancy. This is an observational study, because mothers decided whether or not to smoke during pregnancy; there was no random assignment to smoke or not to smoke. The variables in the dataset are described in the code book at the end of this document.

In 1989, the Surgeon General asserted that mothers who smoke have increased rates of premature delivery (before 270 days) and low birth weights. We will analyze the data to see if there is an association between smoking and pre-term birth, using an indicator variable for pre-term birth. To simplify analyses, we'll compare babies whose mothers smoke to babies whose mothers have never smoked. The data file you have access to has only these people, although there were other types of smokers in the original dataset.

Our questions of interest include the following.

- Do mothers who smoke tend to have higher chances of pre-term birth than mothers who do not smoke? What is a likely range for the difference in odds of pre-term birth for smokers and non-smokers?
- Is there any evidence that the association between smoking and pre-term birth differs by mother's race? If so, characterize those differences.
- Are there other interesting associations with the odds of pre-term birth that are worth mentioning?

Write a report describing your findings. Be sure to include the following in your report: the model you ultimately decided to use, a justification for that model (e.g., why you chose certain transformations and why you decided the final model is reasonable to use based on binned residual diagnostics and ROC curves), the relevant regression output (includes: a table with coefficients and SEs, and p-values or confidence intervals; and somewhere in the text or table the area under the curve for your final model), your interpretation of the results in the context of the questions of interest, and any potential limitations of the analysis.

There are some complexities in this dataset to be aware of. Some variables have missing values. In particular, you will see from the .csv file that the height and weight of the father are missing quite frequently. This is typical in data on births: it is often difficult to get data about the fathers. I recommend that you not consider father's height and weight when modeling. Some of the other variables have a few missing cases here and there. For this analysis, you can drop them from the modeling. This is not the ideal way to handle missing data in an analysis--and we will learn better methods later in the course--but for now it will move the analysis forward. I strongly recommend that you make a data file that has complete observations on every single case for all the variables

you are thinking about including in the model, and run the regression using that file. For example, I posted such a file in the Sakai site that excludes all of the variables on the fathers. You are welcome to use this file, or make your own if you want to use fathers' data.

The file contains an indicator variable for Premature (gestational age < 270 days), which is just a recoding of gestational age; we use that as our outcome variable. The data files also contain two other outcome variables: gestational age and birth weight. Both of these could be affected by smoking, so both are outcomes rather than predictors. It does not make sense scientifically to include one as a predictor of the other; the two variables happen simultaneously and hence are a bivariate outcome. For this analysis, we exclude birth weight from the modeling. Of course, one could do a separate regression for birth weight to see if smoking has an effect on gestational ages. Even better, one could treat birth weight and gestational age as a bivariate outcome and fit a regression model that predicts the bivariate outcome. This is a model we won't have to time to learn about in our course, but come find the instructor if you want to learn more.

The main file also includes information on the number of cigarettes smoked and about timing for mothers who quit smoking. For this analysis you do not have to use those variables, as we just compare smokers and non-smokers.

You can collapse race categories from 0 - 5 into one category for race = white.

Regarding the fathers' data, you might pay attention to correlation among the mothers' and fathers' values. For example, the mothers' and fathers' races might tend to be similar (use a "table" command to see the contingency table of the two races), in which case you have to be concerned about effects of multicollinearity if you want to include both mother's and father's races in the model.

Code Book

Variable	Description
Id	id number
birth	birth date where 1096 = January1, 1961
gestation	length of gestation in days
bwt	birth weight in ounces (999 = unknown)
parity	parity = total number of previous pregnancies, including fetal deaths and still births. 99=unknown
mrace	mother's race or ethnicity 0-5=white 6=mexican 7=black 8=asian 9=mix 99=unknown
mage	mother's age in years at termination of pregnancy
med	mother's education 0 = less than 8th grade 1 = 8th to 12th grade. did not graduate high school 2 = high school graduate, no other schooling 3 = high school graduate + trade school

4 = high school graduate + some college
5 = college graduate
6,7 = trade school but unclear if graduated from high school
9 = unknown

mht mother's height in inches

mpregwt mother's pre-pregnancy weight in pounds

drace father's race or ethnicity
0-5 = white
6 = mexican
7 = black
8 = asian
9 = mix

dage father's age in years at termination of pregnancy

ded father's education
0 = less than 8th grade
1 = 8th to 12th grade. did not graduate high school
2 = high school graduate, no other schooling
3 = high school graduate + trade school
4 = high school graduate + some college
5 = college graduate
6,7 = trade school but unclear if graduated from high school
9 = unknown

dht father's height

dwt father's pre-pregnancy weight in pounds

marital marital status of mother
1 = married
2 = legally separated
3 = divorced
4 = widowed
5 = never married

income family yearly income in 2500 increments. 0 = under 2500, 1 = 2500-4999, ..., 9 = 15000+. 98=unknown, 99=not asked

smoke does mother smoke?
0 = never
1 = smokes now
2 = until preg
3 = once did, not now

time If mother quit, how long ago did she quit?
0 = never smoked,
1 = still smokes,
2 = quit during pregnancy,
3 = up to 1 yr ago,
4 = up to 2 yr ago,

5 = up to 3 yr ago,
6 = up to 4 yr ago,
7 = 5 to 9yr ago,
8 = 10+yr ago,
9 = quit and don't know,
98 = unknown

number number of cigs smoked a day for past and current smokers
0 = never smoked
1 = 1-4
2 = 5-9
3 = 10-14
4 = 15-19
5 = 20-29
6 = 30-39
7 = 40-60
8 = 60+,
9 = smoke but don't know

Premature = 1 if baby born before gestational age of 270, and = 0 otherwise. This is a dichotomized function of the gestational age. We use it as the outcome variable.