Assignment 5
Due: Thursday, March 3

The 1989 Surgeon General's Report claims that

- Mothers who smoke have shorter gestational periods

- The newborns of mothers who smoke are smaller at every gestational age

- Smoking is a more significant determinant of birth weight than any other known risk factor

These claims were based on observational studies which, as we know, are subject to confounding – mother who smoke may differ from nonsmoking mothers in many ways (age, education, income, etc.).

The course website contains data from a comprehensive study of all babies born between 1960 and 1967 at the Kaiser Foundation Hospital in Oakland, California. The course data set contains records only for single births (*i.e.*, not twins) of male babies who survived at least 28 days after birth. It includes the following variables:

- `Gestation`: Length of gestation in days

- `BirthWeight`: Birth weight of the baby, in ounces

- `Parity`: Total number of previous pregnancies for the mother

- `Race`

- `Age`

- `Education`: Mother's education. 0 = Less than 8th grade; 1 = 8th-12th grade (did not graduate); 2 = High school graduate, no other schooling; 3 = High school plus trade school; 4 = High school plus some college; 5 = College graduate; 7 = Trade school, high school graduation unknown

- `MotherHeight`: Mother's height, in inches

- `MotherWeight`: Mother's pre-pregnancy weight, in pounds

- `Marital`: Mother's marital status

- `Income`: Family income in $2,500 increments (0 = under $2,500, 1 = $2,500-$4,999, and so on, up to 9 = over $15,000)

- `Smoke`: Mother's smoking status

In this assignment, your task is to use this data set to address the Surgeon General's claims – do they hold up, or are they the result of confounding? Analyze the data and write a short report (3-5 pages) describing your findings. Include three sections:

- Methods: in which you describe the statistical and methodological aspects of the models that were fit to the data and how you chose a model or models to base your conclusions upon

- Results: in which you present estimates, hypothesis tests, confidence intervals, tables, plots, etc., which describe your findings

- Discussion: in which you briefly recap in everyday language (*i.e.* without statistical jargon) your main conclusions