This midterm consists of two data sets:

1. Cars Data
2. Wages and Hours Data

Each data set is the result of a study conducted by various investigators that are described on the following pages. Along with a description of each study you will find a set of questions posed by each of the investigators prior to collecting the data. These questions are not directly statistical in nature, in that they do not ask for specific analyses to be run. They are, however, the questions that drove the research, and resulted in the data that were collected. Each question can be answered using statistical tools taught in this course up to this point (Chapter 7 in the textbook). *Note that there is not necessarily a single best way to answer each question.* Furthermore, note that this midterm is the statistical equivalent to an essay test, with answers evaluated not only for technical correctness, but also for the motivating reasons for using the statistical methodology – so describe the motivation for your choice of analysis fully. Your job, for each part of the midterm, is to present an answer to each question in the form of an argument that is supported with evidence culled from the statistical analyses. In your mind, consider me as either a journal editor evaluating a manuscript or as a reviewer of a technical report.

Directions:
Your job is to answer the questions given for each data set using statistical techniques you have learned in this class. For **each part of each** question:

- Describe the analysis you have selected to perform (for example, explain why you are using the variables you are using).
- Describe the assumptions of the analysis.
- Report any diagnostic tests of assumptions you may have run.
- Perform the analysis and report the results. Supplying graphs can be helpful to describe your findings.
- Interpret the results.
- Provide an answer to the question using the statistical evidence you have gained via the analysis.
- Describe any limitations of the analysis you performed.
- Remember: direct tabular output from SPSS is not allowed (and will lead to zero points on that section of the analysis); however graphs and figures are encouraged.
- NOTE: Because of logistics, please post your questions to WebCT's discussion board rather than email them to me – this will help others and will allow for communication to flow easier while I am away.

Scoring:
Each data set will represent 1/2 of the total score for the final. For question in the midterm, your score will depend upon these factors:

- Did you choose an analysis that will provide an accurate answer to the question? (10 points)
- Did you explicitly state all assumptions? (10 points)
- Did you perform the appropriate analysis correctly? (20 points)
- Did you check the assumptions? (20 points)
- Were your interpretations consistent with the information the analysis provides? (30 points)
- Were your conclusions well explained (how well did you use the statistical evidence to create an answer to the question)? (10 points)
- BONUS: One percentage point per WebCT discussion post: either a question or an answer (up to 5% of your total grade).

Problem #1
Cars Data

Measurements on 38 1978-79 model automobiles. The gas mileage in miles per gallon as measured by Consumers' Union on a test track. Other values as reported by automobile manufacturer.

Variable Name	Description
Country	Nationality of manufacturer (eg. U.S., Japan)
Car	Car name (Make and model)
MPG	Miles per gallon, a measure of gas mileage
Weight	Weight of the car
Drive_Ratio	Drive ratio of the automobile
Horsepower	Horsepower
Displacement	Displacement of the car (in cubic inches)
Cylinder	Number of cylinders

Researchers at the NTSB want to know:

- How well do weight and drive ratio predict miles per gallon?

- Controlling for weight and drive ratio, how to all other variables correlate with miles per gallon?

- Does a model with all variables predicting miles per gallon do as well at predicting miles per gallon as the model used in part #1?

Problem #2
Wages and Hours Data

The data are from a national sample of 6000 households with a male head and earnings of less than $15,000 annually in 1966. Thirty-nine demographic subgroups were formed for analysis of the relation between average hours worked during the year and average hourly wages and other variables. The study was undertaken in the context of proposals for a guaranteed annual wage (negative income tax).

At issue was the response of labor supply (hours worked) to increasing income and effective hourly wages. If the response was negative, total production in the economy could decline. The idea of a negative income tax as an alternative to traditional welfare programs was proposed independently by Milton Friedman of the University of Chicago and Robert J. Lampman of the University of Wisconsin. The idea was never considered seriously by the Johnson administration, but was considered for a time and then abandoned by the Nixon administration.

Variable Name	Description
HRS	Average hours worked during the year
WAGE	Average hourly wage ($)
ERSP	Average yearly earnings of spouse ($)
ERNO	Average yearly earnings of other family members ($)
NEIN	Average yearly non-earned income
ASSET	Average family asset holdings (Bank account, etc.) ($)
AGE	Average age of respondent
DEP	Average number of dependents
RACE	Percent of white respondents
SCHOOL	Average highest grade of school completed

There are four questions of the study:

- Does the number of average hours (HRS) increase as the average hourly wage (WAGE) of a demographic group increases?

- How do average asset holdings (ASSET), average age (AGE), and average hourly wage (WAGE) predict the average number of hours worked during the year (HRS)?

- Holding average hourly wage (WAGE) constant, what is the correlation between average asset holdings (ASSET) and average number of hours worked during the year (HRS)? How does this compare with the correlation between the two without holding average hours constant?

- Holding average hourly wage (WAGE) constant, what is the correlation between average age (AGE) and average number of hours worked during the year (HRS)? How does this compare with the correlation between the two without holding average hours constant?

Note: the final two questions do not ask for a statistical hypothesis test, just a summary measure. The assumptions, therefore, are very limited.