

## **Incorporating Data Analysis Projects into Introductory Statistics**

- I. Overview
- II. History on Projects
  - A. Project Book
  - B. Term Long Project
  - C. One trick ponies in text books
  - D. Writing as part of General Education Requirement
- III. Goals of the Projects
  - A. Students gaining valuable experience with real data
  - B. Students acquiring communication skills by writing technical reports that summarize results clearly and concisely
  - C. Students learning how to use statistical and word processing software as tools to solve problems and communicate results
  - D. Students acquiring skills in working with others
  - E. Students learning to apply appropriate methodology
- IV. Supporting Documentation
  - A. Context of Cleveland State University
  - B. Group formation Sheet
  - C. Grading Rubric
  - D. Sample Project
  - E. Software Guide
  - F. Summary Lab
- V. Project Topics
  - A. Summary Analysis
  - B. Probability
  - C. Hypothesis Testing
  - D. Regression
- VI. Messy Data Project

## **THINGS TO DO DURING CLASS WHEN FORMING GROUPS**

1. Form groups of four whenever possible.
2. Exchange names, e-mail account (if applicable), and phone numbers with your group members.
3. Choose a group name: the most original and/or wittiest receives a bag of M&M's. A positive attitude regarding mathematics and statistics is encouraged. (e.g. "Math Phobics" is NOT a witty or positive group name.)
4. Expect to spend approximately four hours on each project. Schedule four hours of meeting times for each week during the term. If possible, schedule the meetings in 1-2 hour blocks as opposed to one four hour block. The meetings can be canceled for those weeks when no project is due.
5. Please hand in a copy of the information above (names of group members, e-mail addresses, phone numbers, group name, and meeting times) to your professor.

## **GUIDELINES FOR GROUP PROJECTS**

1. Groups should make every effort to have all members present at each scheduled meeting. It is your responsibility to notify every group member if an emergency arises and you are unable to make a meeting.
2. It is expected that every person in the group is able to justify and explain all parts of the project. On the due date of each project, a random member of the class will be called upon to briefly summarize their group's results. Be prepared for this. In addition, each exam will contain one or more questions from the projects.
3. Each group member will be asked to sign a statement of oath attesting to their involvement in the completion of each project. Under no circumstances should a group member sign this oath if he/she has not participated in the specific project. If a group member is unable to participate in a particular project, please notify the group members and see the instructor prior to the due date of the project for alternatives. With advance notification, it is possible for an individual person to hand in a project.
4. When working in groups, be courteous and respectful of other members. Be sure to listen to all points of view before drawing conclusions. The majority may be wrong.
5. Try to work out all differences and problems within the group. If for some reason your group is not functioning well and unable to resolve the conflict on its own, please set up a meeting time with the professor and the group to discuss your problems.

6. During the completion of the project, it is important that every member participate. It might be helpful to break into pairs and have each pair find the numerical answers to the questions and make the desired graphs. Then check each other to make sure you have the same answers. Be sure to take turns on who runs the mouse. Then you can all talk together on the wording of the report to be handed in as a group. There will be 4 assignments, so appoint a different person to be the team leader for each project. She can make sure each person sees the final version and has opportunity to comment on it.

# Homework Grading Rubric

<b>Organization</b>					
Clarity of Exposition	2	4	6	8	10
Layout	2	4	6	8	10
<b>Mechanics</b>					
Grammar	2	4	6	8	10
Spelling & Punctuation	1	2	3	4	5
<b>Thoroughness</b>	2	4	6	8	10
<b>Professionalism &amp; Style</b>	1	2	3	4	5
<b>Mathematical and Statistical Accuracy (50 points)</b>					
<b>Total Points(100 points maximum)</b>					

We declare that each of the following group members actively contributed to the work handed in. We understand that each group member has the right to veto the signing of another group member who did not contribute to the completion of the assignment.

Group Name:

---

Group Members:

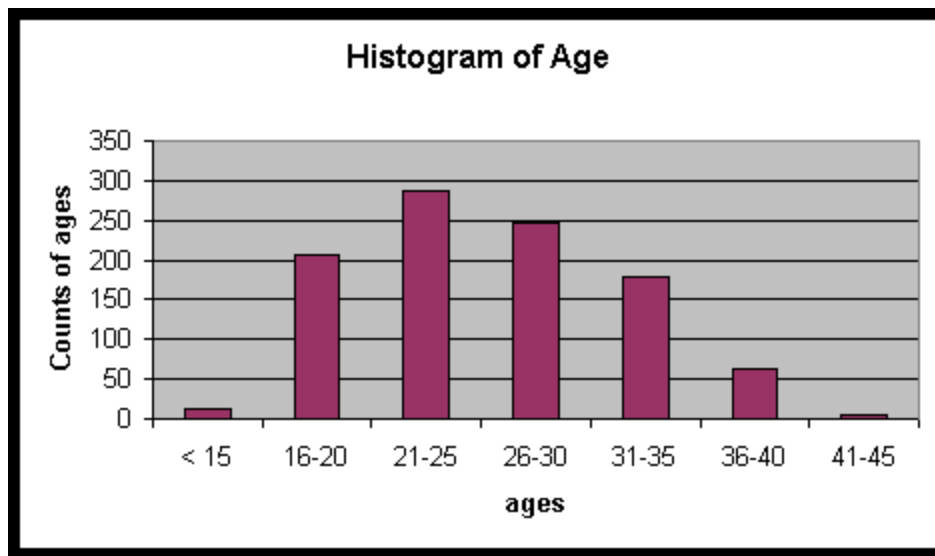

## Project I. THE GREATEST GROUP EVER

The following data analysis report summarizes the information from the data file **ncbirth1450.xls**. This data set contains birth record information from the North Carolina State Center for Health and Environmental Statistics. The data set is a random sample of 1450 births. Birth weight is important because low birth weight has been linked to slower mental and physical development.

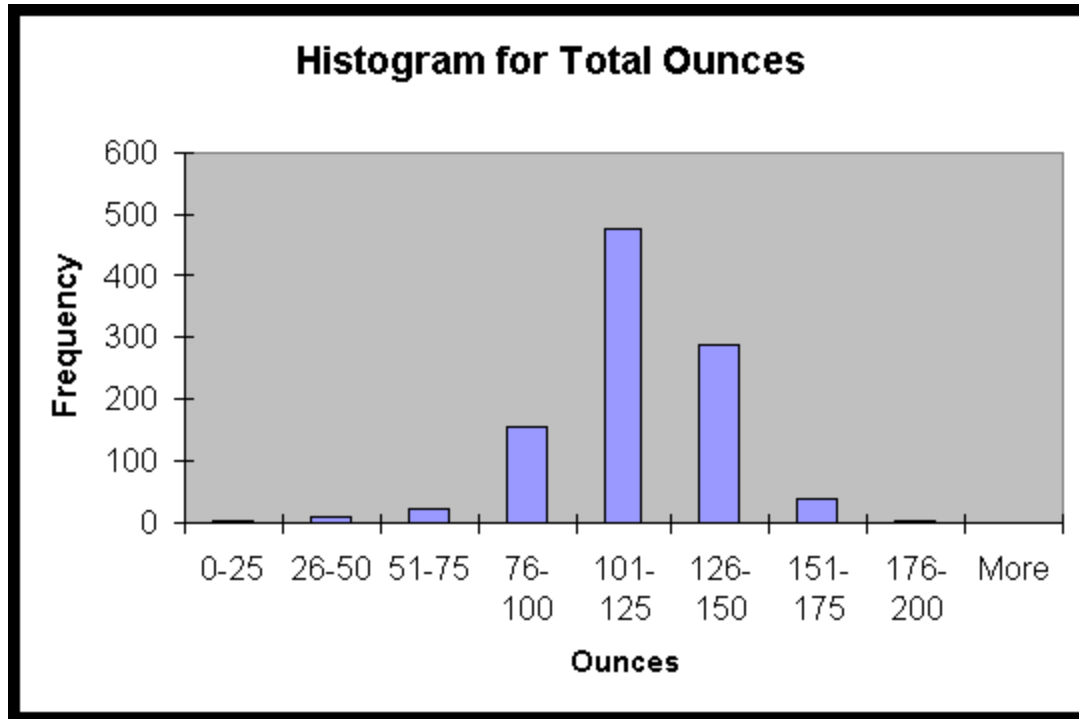
The continuous variables in this study are age and total ounces. A five number summary appears below:

Variable	Mean	Median	SD	Min	Max
age (years)	25.89	25	6.06	13	42
totounc (ounces)	116.13	118	21.57	12	194

Below are the histograms for these variables:



This variable appears to be right skewed so the proper measure of center is the median.



This variable appears slightly left skewed. Since the skewness is only slight the median or mean is a good measure of center.

The discrete variables are **sex**, **educ**, **bwtgroup**, **marital**, and **plural**. Frequency Tables appear for these below:

SEX	Total	Percent
Male 1	496	49.64
Female 2	503	50.35
Grand Total	999	

EDUC	Total	Percent
0	2	0.20
1	1	0.10
2	1	0.10
3	2	0.20
5	3	0.30

6	10	1.00
7	2	0.20
8	29	2.90
9	50	5.01
10	63	6.31
11	63	6.31
12	326	32.63
13	82	8.21
14	110	11.01
15	34	3.40
16	157	15.72
17	62	6.21
Missing	2	0.2002
Grand Total	999	

<b>BWTGROUP</b> (grams)	Total	Percent
<500 0	3	0.30
501-1000 1	4	0.40
1001-1500 2	7	0.70
1501-2000 3	14	1.40
2001-2500 4	58	5.81
2501-3000 5	178	17.82
3001-3500 6	357	35.74
3501-4000 7	286	28.63
4001-4500 8	80	8.10
> 4501 9	11	1.10

Missing	1	0.10
Grand Total	999	

MARITAL	Total	Percent
Married 1	671	67.17
Unmarried 2	326	32.63
Missing	2	0.20
Grand Total	999	

PLURAL	Total	Percent
Single 1	975	97.5976
Twins 2	23	2.302302
Triplets 3	1	0.1001
Grand Total	999	

The frequency tables for the three variables low, smoke, and drankalc appear below:

LOW	Total	Percent
0	912	91.29
1	86	8.61
Missing	1	0.10
Grand Total	999	

SMOKE	Total	Percent
Non-smoker	824	82.48
Smoker	173	17.32
Missing	2	0.20
Grand Total	999	

DRANKALC	Total	Percent
0	984	98.50

1	13	1.30
Missing	2	0.20
Grand Total	999	

Yes, I think having a 10lb baby is unlikely – it is huge! That poor mother had to carry that heavy-sucker for a long time. I would tell my friend if she gained 30 pounds during pregnancy not to worry because that doesn't sound too bad to me. I think it should not matter if a woman smokes during pregnancy. I smoked during my own pregnancy and my son weighed over eight pounds. Of course, I did not drink during pregnancy. I would ask the mother about drug use, nutrition, and the helpfulness of the father during pregnancy.

## Summary Lab/SPSS Introduction

### I. Enter Data

Enter into three columns average monthly temperature data from Raleigh, NC and San Francisco, CA.

Month	Raleigh	San Francisco
January	39	49
February	42	52
March	50	53
April	59	56
May	67	58
June	74	62
July	78	63
August	77	64
September	71	65
October	60	61
November	51	55
December	43	49

II. Use the **Variable View** Tab to enter names for the variables, change the decimals shown, and give it a label.

III. Find the Mean, Median, Minimum, Maximum, Standard Deviation of both cities. **[Analyze\Descriptive Statistics\Explore]**

	Mean	Median	Min	Max	SD
Raleigh					
San Francisco					

IV. Now go to the web page <http://academic.csuohio.edu/holcombj> and click on the link for MTH 147. On that page, download the dataset **studsurvey147.sav**. Add Labels and codes by clicking the **Variable View** tab.

V. Open the data in SPSS and create the frequency tables for the categorical variables. **[Analyze\Descriptive Statistics\Frequencies]** Fill in the missing numbers below.

#### CAFFEINE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	19		57.6	57.6
	Yes		42.4	42.4	100.0
Total		33	100.0	100.0	

**GENDER**

		Percent	Valid Percent	Cumulative Percent
Valid	Male	69.7	69.7	69.7
	Female	30.3	30.3	100.0
	Total	100.0	100.0	

**EXERCISE**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Rarely	12		36.4	36.4
	Moderately	13		39.4	75.8
	Extensively	8		24.2	100.0
	Total	33	100.0	100.0	

**22 or Under**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	22 or less	21		63.6	63.6
	over 22	12		36.4	100.0
	Total	33	100.0	100.0	

**Have you seen a movie since Aug. 15**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No		63.6	63.6	63.6
	Yes			36.4	100.0
	Total	33	100.0	100.0	

**Major**

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid art	1	3.0	3.0	3.0
biology		18.2	18.2	21.2
biotech	1	3.0	3.0	24.2
comm	2		6.1	30.3
criminal	1	3.0	3.0	33.3
educatioin	1	3.0	3.0	36.4
education			27.3	63.6
env. science	1	3.0	3.0	66.7
health	1	3.0	3.0	69.7
liberal arts	1	3.0	3.0	72.7
nursing		3.0	3.0	75.8
political	1	3.0	3.0	78.8
psychology	1	3.0	3.0	81.8
sociology	3		9.1	90.9
sports	2	6.1	6.1	97.0
undecided	1	3.0	3.0	100.0
Total	33	100.0	100.0	

VI. Now summarize the continuous variables [Analyze\Descriptive Statistics\Explore]:

	Mean	Median	Min	Max	SD
<b>Resting</b>					
<b>Active</b>					
<b>Height</b>					
<b>Foot Length</b>					
<b>Armspan</b>					
<b>Work</b>					

## Project #1: Data Summary

The data which is stored in the file **ncbirth1450.sav** is a random sample of 1450 birth records taken by the North Carolina State Center for Health and Environmental Statistics in 2001. Of particular interest will be incidents of low infant birth weight. Low birth weight is commonly defined as less than 2500 grams. Over the course of the semester we will investigate the relationship of several variables with low birth weight and each other. The goal of this assignment will be to summarize the variables in this data set both graphically and numerically. The variables in this study are:

Variable Label	Description
<b>plurality</b>	Number of children born of the pregnancy
<b>sex</b>	Sex of child (1=Male, 2=Female)
<b>mage</b>	Age of mother (years)
<b>weeks</b>	Completed Weeks of Gestation (weeks)
<b>marital</b>	Marital status (1=married, 2=not married)
<b>racemom</b>	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
<b>hispmom</b>	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
<b>gained</b>	Weight gained during pregnancy (pounds)
<b>smoke</b>	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
<b>drink</b>	0=mother did not consume alcohol during pregnancy 1=mother did consume alcohol during pregnancy
<b>tounces</b>	Weight of child (ounces)
<b>tgrams</b>	Weight of child (grams)
<b>low</b>	0=infant was not low birth weight 1=infant was low birth weight
<b>Premie</b>	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

The goal of this assignment is to obtain summary statistics for the variables in the data set. This is an important activity of most statistical studies. In your report, clearly label all tables and when appropriate give the units of measure. The components of the assignment are given below. Be sure your presentation is clear and organized. The use of tables is required.

First identify the categorical variables and the continuous variables. For the categorical variables construct a frequency table that gives the counts and percentages of observations in each category. For the continuous variables, create a histogram, comment on the shape of the histogram, and determine if the mean or median is the most appropriate measure of center. For each continuous variable, create a table that gives the mean, median, standard deviation, minimum and maximum value.

Suppose a friend of yours has given birth to a 10 pound 3 ounce baby. Is this unusual? Why or why not? In writing explain your reasoning. Another friend of yours had a baby and gained approximately 30 pounds during her pregnancy. Explain to her why she should not be too depressed over this occurrence. Comment on the percentage of women who reported smoking and drinking during pregnancy.

Lastly, propose three other variables you would like to investigate in regard to weight of the mother. Give three explicit questions you would ask the mother prior to delivery and explain why you want to know that information.

## Project #2: Probability Summary

Recall the data on birth weight gathered from the North Carolina State Center for Health and Environmental Statistics. Use the data set of 1450 observations **ncbirth1450.sav**. Recall the variables the variables examined are:

Variable Label	Description
<b>plurality</b>	Number of children born of the pregnancy
<b>sex</b>	Sex of child (1=Male, 2=Female)
<b>mage</b>	Age of mother (years)
<b>weeks</b>	Completed Weeks of Gestation (weeks)
<b>marital</b>	Marital status (1=married, 2=not married)
<b>racemom</b>	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
<b>hispmom</b>	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
<b>gained</b>	Weight gained during pregnancy (pounds)
<b>smoke</b>	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
<b>drink</b>	0=mother did not consume alcohol during pregnancy 1=mother did consume alcohol during pregnancy
<b>tounces</b>	Weight of child (ounces)
<b>tgrams</b>	Weight of child (grams)
<b>low</b>	0=infant was not low birth weight 1=infant was low birth weight
<b>premie</b>	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

Pregnant women are encouraged by physicians to refrain from such activities as smoking. It is believed that women who engage in these activities are at higher risk for having a low birth weight child. This assignment will have you investigate whether this actually turns out to be the case with the data above.

Create a contingency table of **smoke** vs. **low**. Give an estimate of the probability of a low birth weight child. Give an estimate of the probability that the child's mother smokes. Estimate the probability of low birth weight given the mother smokes. Estimate the probability of low birth weight given the mother does not smoke. Calculate the prevalence ratio of low birth weight vs. smoking status and interpret its meaning. Take into account the missing values when making your calculations.

Obviously, premature babies are more likely to be low birth weight, but how obvious? Create a table of **premie** vs. **low**. Calculate the probability a baby is born premature. Calculate the probability the baby is low birth weight given it is premature. Calculate the probability the baby is low birth weight given it is not premature. Calculate the prevalence ratio of low birth weight if the baby is premature. Interpret its meaning.

Create a table of **smoke** vs. **premie**. Using **smoke** as the risk factor, calculate the probability of a premature baby given the mother smoked, the probability of a premature baby given the mother did not smoke, and the prevalence ratio of a premature baby with smoking as the risk factor. Interpret your results.

Create a table of **marital** vs. **low** and **marital** vs. **premie**. For each table use **marital=0** as the risk factor and find the prevalence ratio of low birth weight and prevalence ratio of premature birth. Interpret your results.

Dr. Holcomb's mother is quite angry at the above analysis. She smoked during her two pregnancies and both of her children weighed above 6 pounds (ie they were not low birth weight). She thinks her son has wasted his life studying a useless subject like statistics because she smoked and her kids turned out fine. Explain in words the error in her thinking.

After doing the analysis above, a MTH 147 student proclaims that being unmarried causes premature birth and low birth weight. Is this logic correct? Explain in words why or why not.

## Project #3 Hypothesis Testing Project

For this project, we will use a subset of the North Carolina birth data set. The data set **ncbirth200.sav** is a random sample of 200 births from the data set **ncbirth1450.sav**. When doing this assignment, make sure you are working with the data set with only 200 observations! In this assignment you will test hypotheses relating to **mage**, **weeks**, **tounces**, **low**, and **smoke**.

Variable Label	Description
<b>plurality</b>	Number of children born of the pregnancy
<b>sex</b>	Sex of child (1=Male, 2=Female)
<b>mage</b>	Age of mother (years)
<b>weeks</b>	Completed Weeks of Gestation (weeks)
<b>marital</b>	Marital status (1=married, 2=not married)
<b>racemom</b>	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
<b>hispmom</b>	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
<b>gained</b>	Weight gained during pregnancy (pounds)
<b>smoke</b>	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
<b>drink</b>	0=mother did not consume alcohol during pregnancy 1=mother did consume alcohol during pregnancy
<b>tounces</b>	Weight of child (ounces)
<b>tgrams</b>	Weight of child (grams)
<b>low</b>	0=infant was not low birth weight 1=infant was low birth weight
<b>Premie</b>	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

Begin the assignment by proving a frequency table for the percentage of low birth weights and a frequency table for the percentage of smokers. Create a summary table (mean, median, standard deviation, minimum, maximum) for the continuous variables of **mage**, **weeks**, and **tounces**, **gained** broken down by smoking status.

With the information that you gather from this summary, test the following (you will need to do the tests of a proportion by hand, but the test of a mean can be done using the computer):

- a. Determine if there is sufficient evidence to conclude the mean age of mothers giving birth in North Carolina is over 25 years of age at the 0.05 level of significance.
- b. Determine if there is sufficient evidence to conclude the mean weeks of gestation of mothers giving birth in North Carolina is below 39 weeks.
- c. Determine if there is sufficient evidence to conclude that the mean weight of babies born to mothers in North Carolina is above 7 lbs. (Note that there are 16 ounces in a pound.)
- d. Determine if there is sufficient evidence to conclude the percentage of low birth weight children in North Carolina is above 6%.
- e. Determine if there is sufficient evidence to conclude the percentage of mothers who smoke in North Carolina is above 10%.
- f. Construct a side-by-side boxplot for **tounces** for smokers and non-smokers. Comment on whether you believe you will reject or fail to reject the null hypothesis. Determine if there is sufficient evidence to conclude the mean **tounces** of smoking mothers is lower than the mean birth weight for non-smoking mothers.

For each of the tests above, in your report, be sure to

1. Clearly state a null and alternative hypothesis
2. Give the value of the test statistic
3. Report the P-value
4. Clearly state your conclusion (i.e. 'Reject the Null' is not sufficient)

For d and e above, be sure to check the assumptions associated with a test of a proportion.

Lastly, propose and conduct your own test of hypotheses. You can test a single mean, a single proportion or compare two means for two independent groups. Make sure your test follows the four steps above.

## Project #4: Simple Linear Regression

For this assignment use the data set **ncbirth200.sav**. Recall the variables from the North Carolina birth data set are:

The variables examined are:

Variable Label	Description
<b>plurality</b>	Number of children born of the pregnancy
<b>sex</b>	Sex of child (1=Male, 2=Female)
<b>mage</b>	Age of mother (years)
<b>weeks</b>	Completed Weeks of Gestation (weeks)
<b>marital</b>	Marital status (1=married, 2=not married)
<b>racemom</b>	Race of Mother (0=Other Non-white, 1=White, 2=Black 3=American Indian, 4=Chinese, 5=Japanese, 6=Hawaiian, 7=Filipino, 8=Other Asian or Pacific Islander)
<b>hispmom</b>	Mother of Hispanic origin (C=Cuban, M=Mexican, N=Non-Hispanic, O=Other and Unknown Hispanic, P=Puerto Rican, S=Central/South American, U=Not Classifiable)
<b>gained</b>	Weight gained during pregnancy (pounds)
<b>smoke</b>	0=mother did not smoke during pregnancy 1=mother did smoke during pregnancy
<b>drink</b>	0=mother did not consume alcohol during pregnancy 1=mother did consume alcohol during pregnancy
<b>tounces</b>	Weight of child (ounces)
<b>tgrams</b>	Weight of child (grams)
<b>low</b>	0=infant was not low birth weight 1=infant was low birth weight
<b>Premie</b>	0=infant was not premature 1=infant was premature premature defined at 36 weeks or sooner

Answer the following for the variables **tounces** (dependent variable) and **mage** (independent variable).

- a. Make a scatterplot of this data. Fit the regression line. Report the parameter estimates (the estimates of the intercept and slope).
- b. Is **mage** useful in predicating **tounces**? Why? Report the level of significance (P-value).
- c. What percentage of the variation in **tounces** is explained by **mage**? Is that high or low?
- d. What is the predicted value for **tounces** when **mage** is 35? What if **mage** is 17?
- e. Make a residual plot. Comment on the fit of the model.

Answer the following for the variables **tounces** (dependent variable) and **weeks** (independent variable).

- a. Make a scatterplot of this data. Fit the regression line. Report the parameter estimates.
- b. Is **weeks** useful in predicating **tounces**? Why? Report the P-value.
- c. What percentage of the variation in **tounces** is explained by **weeks**? Is that high or low?
- d. What is the predicted value for **tounces** when **weeks** is 35? What if **weeks** is 40?
- e. Make a residual plot. Comment on the fit of the model.

## Assignment I, MTH 567, Fall 2003

In the data set **calcium.sas7bdat**, there are three discrete variables, **sex**, **lab**, and **agegroup**. The coding is as follows:

Sex	1=Male; 2=Female
Lab	1=Metpath; 2=Deyor; 3=St. Elizabeth's; 4=CB Rouche; 5=YOH; 6=Horizon
Agegroup	1=65-69; 2=70-74; 3=75-79; 4=80-84; 5=85-89 years

The other variables of **age**, **alkphos** - alkaline phosphatase (IU/L), **ca** - raw calcium (mg/dL), **iphos** - inorganic phosphorus (mg/dL), **cammol** - calcium (mmol/L), and **phosmmol** - phosphorus (mmol/L), are continuous.

1. The first task of the assignment is to check the validity of the data. Determine if this is a "messy" data set with variable values that appear incorrect. Attempt to recover the correct values by looking up the true values from the actual data records. Copies of these can be found on <http://academic.csuohio.edu/holcombj/outlier/bigtable.htm>
2. Once the data is "clean", perform a summary analysis of the three discrete variables. For the variables **alkphos**, **ca** and **iphos**, report the mean, median, standard deviation, min and max broken down by **sex**. Also summarize the variables **alkphos**, **ca** and **iphos** in a similar way with the factor variable as **lab**.
3. Then Construct side by side box plots of the variables **alkphos**, **ca**, and **iphos** with the factor variable as **sex**. Then construct side by side box plots of the **alkphos**, **ca**, and **iphos** continuous variables with the factor variable as **lab**.
4. Do you believe a significant difference exists in **alkphos**, **ca**, or **iphos** levels with respect to **sex**? Why or why not? Do you believe a significant difference exists in **alkphos**, **ca**, or **iphos** levels with respect to **lab**? Why or why not?
5. Suppose Mr and Mrs. Contrarian are married and Mrs. Contrarian has lower calcium than Mr. Contrarian. She refuses to believe the results of the study that men tend to have lower calcium than women because she has lower calcium than her husband. Using your results to question #3, explain to Mrs. Contrarian the flaw in her thinking.
6. One the objectives of the research for the study from which this data came was to argue for separate reference ranges for men and women who are over age 64. Looking at the different results for raw calcium (**ca**) between the different labs, explain why a single reference range is so difficult to establish.