Draft Lecture Notes for Methods of Applied Statistics (STA442H)

Jerry Brunner

January 2, 2004

Chapter 1

Introduction

This course is about using statistical methods to draw conclusions from real data. It is deliberately non-mathematical, relying on translations of statistical theory into English. For the most part, formulas are avoided. This involves some loss of precision, it also makes the course accessible to students from non-statistical disciplines (particularly graduate students and advanced undergraduates on their way to graduate school) who need to use statistics in their research. Even for students with strong training in theoretical statistics, the use of plain English can help reveal the connections between theory and applications, while also suggesting a useful way to communicate with non-statisticians.

We will avoid mathematics, but we will not avoid computers. Learning to apply statistical methods to real data involves actually doing it, and the use of software is not optional. Furthermore, we will *not* employ "user-friendly" menu-driven statistical programs. Why?

- It's just too easy to poke around in the menus trying different things, produce some results that seem reasonable, and then two weeks later be unable to say exactly what one did.

- Real data sets tend to be large and complex, and most statistical analyses involve a sizeable number of operations. If you discover a tiny mistake after you produce your results, you don't want to go back and repeat two hours of menu selections and mouse clicks, with one tiny variation.

- If you need to analyze a data set that is similar to one you have analyzed in the past, it's a lot easier to edit a program than to remember a collection of menu selections from last year.

Don't worry! The word "program" does *not* mean we are going to write programs in some true programming language like C or Java. We'll use statistical software in which most of the actual statistical procedures have already been written by experts; usually, all we have to do is invoke them by using high-level commands.

The statistical packages we will use in this course are `SAS` and `S`. These packages are command-oriented rather than menu-oriented, and are very powerful. They are industrial strength tools, and will be illustrated in an industrial strength environment — `unix`. This is mostly for local convenience. There are Windows versions of both `SAS` and `S` that work just as well as the unix versions, except for very big jobs.

Applied Statistics really refers to two related enterprises. The first might be more accurately termed "Applications of Statistics," and consists of the appropriate application of standard general techniques. The second enterprise is the development of specialized techniques that are designed specifically for the data at hand. The difference is like buying your clothes from Walmart versus sewing them yourself (or going to a tailor). In this course, we will do both. We'll maintain the non-mathematical nature of the course in the second half by substituting computing power and random number generation for statistical theory.

1.1 Vocabulary of data analysis

We start with a **data file**. Think of it as a rectangular array of numbers, with the rows representing **cases** (units of analysis, observations, subjects, replicates) and the columns representing **variables** (pieces of information available for each case).

- A physical data file might have several lines of data per case, but you can imagine them listed on a single long line.
- Data that are *not* available for a particular case (for example because a subject fails to answer a question, or because a piece of measuring equipment breaks down) will be represented by missing value codes. Missing value codes allow observations with missing information to be automatically excluded from a computation.
- Variables can be **quantitative** (representing amount of something) or **categorical**. In the latter case the "numbers" are codes representing category membership. Categories may be **ordered** (small vs. medium vs. large) or **unordered** (green vs. blue vs. yellow). When a quantitative variable reflects measurement on a scale capable of very fine gradation, it is sometimes described as **continuous**. Some statistical texts use the term **qualitative** to mean categorical. When an anthropologist uses the word "qualitative," however, it usually means "non-quantitative."

Another very important way to classify variables is

Independent Variable (IV): Predictor = X (actually $X_i, i = 1, \ldots, n$)

Dependent Variable (DV): Predicted = Y (actually $Y_i, i = 1, \ldots, n$)

Example: X = weight of car in kilograms, Y = fuel efficiency in litres per kilometer

Sample Question 1.1.1 *Why isn't it the other way around?*

Answer to Sample Question 1.1.1 *Since weight of a car is a factor that probably influences fuel efficiency, it's more natural to think of predicting fuel efficiency from weight.*

The general principle is that if it's more natural to think of predicting A from B, then A is the dependent variable and B is the independent variable. This will usually be the case when B is thought to cause or influence A. Sometimes it can go either way or it's not clear. Usually it's easy to decide.

Sample Question 1.1.2 *Is it possible for a variable to be both quantitative and categorical? Answer Yes or No, and either give an example or explain why not.*

Answer to Sample Question 1.1.2 *Yes. For example, the number of cars owned by a person or family.*

In some fields, you may hear about **nominal, ordinal, interval** and **ratio** variables, or variables measured using "scales of measurement" with those names. Ratio means the scale of measurement has a true zero point, so that a value of 4 represents twice as much as 2. An interval scale means that the difference (interval) between 3 and 4 means the same thing as the difference between 9 and 10, but zero does not necessarily mean absence of the thing being measured. The usual examples are shoe size and ring size. In ordinal measurement, all you can tell is that 6 is less than 7, not how much more. Measurement on a nominal scale consists of the assignment of unordered categories. For example, citizenship is measured on a nominal scale.

It is usually claimed that one should calculate means (and therefore, for example, do multiple regression) only with interval and ratio data; it's usually acknowledged that people do it all the time with ordinal data, but they really shouldn't. And it is obviously crazy to calculate a mean on numbers representing unordered categories. Or is it?

Sample Question 1.1.3 *Give an example in which it's meaningful to calculate the mean of a variable measured on a nominal scale.*

Answer to Sample Question 1.1.3 *Code males as zero and females as one. The mean is the proportion of females.*

It's not obvious, but actually all this talk about what you should and shouldn't do with data measured on these scales does not have anything to do with *statistical* assumptions. That is, it's not about the mathematical details of any statistical model. Rather, it's a set of guidelines for what statistical model one ought to adopt. Are the guidelines reasonable? It's better to postpone further discussion until after we have seen some details of multiple regression.

1.2 Statistical significance

We will often pretend that our data represent a **random sample** from some **population**. We will carry out formal procedures for making inferences about this (usually fictitious) population, and then use them as a basis for drawing conclusions about the data.

Why do we do all this pretending? As a formal way of filtering out things that happen just by coincidence. The human brain is organized to find *meaning* in what it perceives, and it will find apparent meaning even in a sequence of random numbers. The main purpose of testing for statistical significance is to protect Science against this. Even when the data do not fully satisfy the assumptions of the statistical procedure being used (for example, the data are not really a random sample) significance testing can be a useful as a way of restraining scientists from filling the scientific literature with random garbage. This is such an important goal that we will spend almost the entire course on significance testing.

1.2.1 Definitions

Numbers that can be calculated from sample data are called **statistics**. Numbers that could be calculated if we knew the whole population are called **parameters**. Usually parameters are represented by Greek letters such as α, β and γ, while statistics are represented by ordinary letters such as a, b, c. Statistical inference consists of making decisions about parameters based on the values of statistics.

The **distribution** of a variable corresponds roughly to a histogram of the values of the variable. In a large population for a variable taking on many values, such a histogram will be indistinguishable from a smooth curve.

For each value x of the independent variable X, in principle there is a separate distribution of the dependent variable Y. This is called the **conditional distribution** of Y given $X = x$.

We will say that the independent and dependent variable are **unrelated** if the *conditional distribution of the dependent variable in the population is identical for each value of the independent variable*. That is, the histogram of the dependent variable does not depend on the value of the independent variable. If the distribution of the dependent variable does depend on the value of the independent variable, we will describe the two variables as **related**.

Most research questions involve more than one independent variable. It is also common to have more than one dependent variable. When there is one dependent variable, the analysis is called **univariate**. When more than one dependent variable is being considered simultaneously, the analysis is called **multivariate**.

Sample Question 1.2.1 *Give an example of a study with two categorical independent variables, one quantitative independent variable, and two quantitative dependent variables.*

Answer to Sample Question 1.2.1 *In a study of success in university, the subjects are first-year university students. The categorical independent variables are Sex and Immigration Status (Citizen, Permanent Resident or Visa), and the quantitative independent variable is family income. The dependent variables are cumulative Grade Point Average at the end of first year, and number of credits completed in first year.*

Many problems in data analysis reduce to asking whether one or more variables are related – not in the actual data, but in some hypothetical population from which the data are assumed to have been sampled. The reasoning goes like this. Suppose that the independent and dependent variables are actually unrelated *in the population*. If this is true, what is the probability of obtaining a *sample* relationship between the variables that is as strong or stronger than the one we have observed? If the probability is small (say, $p < 0.05$), then we describe the sample relationship as **statistically significant**, and it is socially acceptable to discuss the results. In particular, there is some chance of having the results taken seriously enough to publish in a scientific journal.

Here is another way to talk about p-values and significance testing. *The p-value is the probability of getting our results (or better) just by chance.* If p is small enough (we will use) then the data are very unlikely to have arisen by chance, assuming there is really no relationship between the independent variable and the dependent variable in the population. In this case we will conclude there is a relationship between the independent and dependent, and we will say our results are "statistically significant."

If $p > .05$, we will not conclude anything. All we can say is that there is no evidence of a relationship between the independent variable and the dependent variable.

For those who like precision, the formal definition is this. The p-value is the minimum significance level α at which the null hypothesis (of no relationship between IV and DV in the population) can be rejected.

1.2.2 Standard elementary significance tests

We will now consider some of the most common elementary statistical methods. For each one, you should be able to answer the following questions.

1. Make up your own original example of a study in which the technique could be used.
2. In your example, what is the independent variable (or variables)?
3. In your example, what is the dependent variable (or variables)?
4. Indicate how the data file would be set up.

Independent observations One assumption shared by most standard methods is that of *"independent observations."* The meaning of the assumption is this. Observations 13 and 14 are independent if and only if the conditional distribution of observation 14 given observation 13 is the same for each possible value observation 13. For example if the observations are temperatures on consecutive days, this would not hold. If the dependent variable is score on a homework assignment and students copy from each other, the observations will not be independent.

When significance testing is carried out under the assumption that observations are independent but really they are not, results that are actually due to chance will often be detected as significant with probability considerably greater than 0.05. This is sometimes called the problem of *inflated n*. In other words, you are pretending you have more separate pieces of information than you really do. When observations cannot safely be assumed independent, this should be taken into account in the statistical analysis. We will return to this point again and again.

Independent (two-sample) t-test

This is a test for whether the means of two independent groups are different. Assumptions are independent observations, normality within groups, equal variances. For large samples normality does not matter. For large samples with nearly equal sample sizes, equal variance assumption does not matter. The assumption of independent observations is always important.

Sample Question 1.2.2 *Make up your own original example of a study in which a two-sample t-test could be used.*

Answer to Sample Question 1.2.2 *An agricultural scientist is interested in comparing two types of fertilizer for potatoes. Fifteen small plots of ground receive fertilizer A and fifteen receive fertilizer B. Crop yield for each plot in pounds of potatoes harvested is recorded.*

Sample Question 1.2.3 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.3 *Fertilizer, a binary variable taking the values A and B.*

Sample Question 1.2.4 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.4 *Crop yield in pounds.*

Sample Question 1.2.5 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.5

A	13.1
A	11.3
⋮	⋮
B	12.2
⋮	⋮

Matched (paired) t-test

Again comparing two means, but from paired observations. Pairs of observations come from the same case (subject, unit of analysis), and presumably are non-independent. Again, the data from a given pair are not really separate pieces of information, and if you pretend they are, then you are pretending to have more accurate estimation of population parameters — and a more sensitive test — than you really do. The probability of getting results that are statistically significant will be greater than 0.05, even if nothing is going on.

In a matched t-test, this problem is taken care of by computing a difference for each pair, reducing the volume of data (and the apparent sample size) by half. This is our first example of a *repeated measures* analysis. Here is a general definition. We will say that there are **repeated measures** on an independent variable if a case (unit of analysis, subject, participant in the study) contributes a value of the dependent variable for each value of the independent variable in question. A variable on which there are repeated measures is sometimes called a **within-subjects** variable. When this language is being spoken, variables on which there are not repeated measures are called **between-subjects**.

The assumptions of the matched t-test are that the differences represent independent observations from a normal population. For large samples, normality does not matter. The assumption that different cases represent independent observations is always important.

Sample Question 1.2.6 *Make up your own original example of a study in which a matched t-test could be used.*

Answer to Sample Question 1.2.6 *Before and after a 6-week treatment, participants in a quit-smoking program were asked "On the average, how many cigarettes do you smoke each day?"*

Sample Question 1.2.7 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.7 *Presence versus absence of the program, a binary variable taking the values "Absent" or "Present" (or maybe "Before" and "After"). We can say there are* repeated measures *on this factor, or that it is a* within-subjects *factor.*

Sample Question 1.2.8 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.8 *Reported number of cigarettes smoked per day.*

Sample Question 1.2.9 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.9 *The first column is "Before," and the second column is "After."*

$$\begin{array}{cc} 22 & 18 \\ 40 & 34 \\ 20 & 10 \\ \vdots & \vdots \end{array}$$

One-way Analysis of Variance

Extension of the independent t-test to two or more groups. Same assumptions, everything. $F = t^2$ for two groups.

Sample Question 1.2.10 *Make up your own original example of a study in which a one-way analysis of variance could be used.*

Answer to Sample Question 1.2.10 *Eighty branches of a large bank were chosen to participate in a study of the effect of music on tellers' work behaviour. Twenty branches were randomly assigned to each of the following 4 conditions. 1=No music, 2=Elevator music, 3=Rap music, 4=Individual choice (headphones). Average customer satisfaction and worker satisfaction were assessed for each bank branch, using a standard questionnaire.*

Sample Question 1.2.11 *In your example, what are the cases?*

Answer to Sample Question 1.2.11 *Branches, not people answering the questionnaire.*

Sample Question 1.2.12 *Why do it that way?*

Answer to Sample Question 1.2.12 *To avoid serious potential problems with independent observations within branches. The group of interacting people within social setting is the natural unit of analysis, like an organism.*

Sample Question 1.2.13 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.13 *Type of music, a categorical variable taking on 4 values.*

Sample Question 1.2.14 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.14 *There are 2 dependent variables, average customer satisfaction and average worker satisfaction. If they were analyzed simultaneously the analysis would be multivariate (and not elementary).*

Sample Question 1.2.15 *Indicate how the data file might be set up.*

Answer to Sample Question 1.2.15 *The columns correspond to Branch, Type of Music, Customer Satisfaction and Worker Satisfaction*

$$\begin{array}{cccc} 1 & 2 & 4.75 & 5.31 \\ 2 & 4 & 2.91 & 6.82 \\ \vdots & \vdots & \vdots & \vdots \\ 80 & 2 & 5.12 & 4.06 \end{array}$$

Sample Question 1.2.16 *How could this be made into a repeated measures study?*

Answer to Sample Question 1.2.16 *Let each branch experience each of the 4 music conditions in a random order (or better, use only 72 branches, with 3 branches receiving each of the 24 orders). There would then be 16 pieces of data for each bank.*

Including all orders of presentation in each experimental condition is an example of **counterbalancing** — that is, presenting stimuli in such a way that order of presentation is unrelated to experimental condition. That way, the effects of the treatments are not confused with fatigue or practice effects (on the part of the experimenter as well as the subjects). In counterbalancing, it is often not feasible to include *all* possible orders of presentation it each experimental condition, because sometimes there are two many. The point is that order of presentation has to be unrelated to any manipulated independent variable.

Two (and higher) way Analysis of Variance

Extension of One-Way ANOVA to allow assessment of the joint relationship of several categorical independent variables to one quantitative dependent variable that is assumed normal within treatment combinations. Tests for interactions between IVs are possible. An interaction means that the relationship of one independent variable to the dependent variable *depends* on the value of another independent variable. More on this later.

Crosstabs and chisquared tests

Cross-tabulations (Crosstabs) are joint frequency distribution of two categorical variables. One can be considered an IV, the other a DV if you like. In any case (even when the IV is manipulated in a true experimental study) we will test for significance using the *chi-squared test of independence*. Assumption is independent observations are drawn from a multinomial distribution. Violation of the independence assumption is common and very serious.

Sample Question 1.2.17 *Make up your own original example of a study in which this technique could be used.*

Answer to Sample Question 1.2.17 *For each of the prisoners in a Toronto jail, record the race of the offender and the race of the victim. This is illegal; you could go to jail for publishing the results. It's totally unclear which is the IV and which is the DV, so I'll make up another example.*

For each of the graduating students from a university, record main field of study and and gender of the student (male or female).

Sample Question 1.2.18 *In your example, what is the independent variable (or variables)?*

Answer to Sample Question 1.2.18 *Gender*

Sample Question 1.2.19 *In your example, what is the dependent variable (or variables)?*

Answer to Sample Question 1.2.19 *Main field of study (many numeric codes).*

Sample Question 1.2.20 *Indicate how the data file would be set up.*

Answer to Sample Question 1.2.20 *The first column is Gender (0=Male, 1=F). The second column is Field.*

$$\begin{array}{cc} 1 & 2 \\ 0 & 14 \\ 0 & 9 \\ \vdots & \vdots \end{array}$$

Correlation and Simple Regression

Correlation Start with a **scatterplot** showing the association between two (quantitative, usually continuous) variables. A scatterplot is a set of Cartesian coordinates with a dot or other symbol showing the location of each (x, y) pair. If one of the variables is clearly the independent variable, it's traditional to put it on the x axis. There are n points on the scatterplot, where n is the number of cases in the data file.

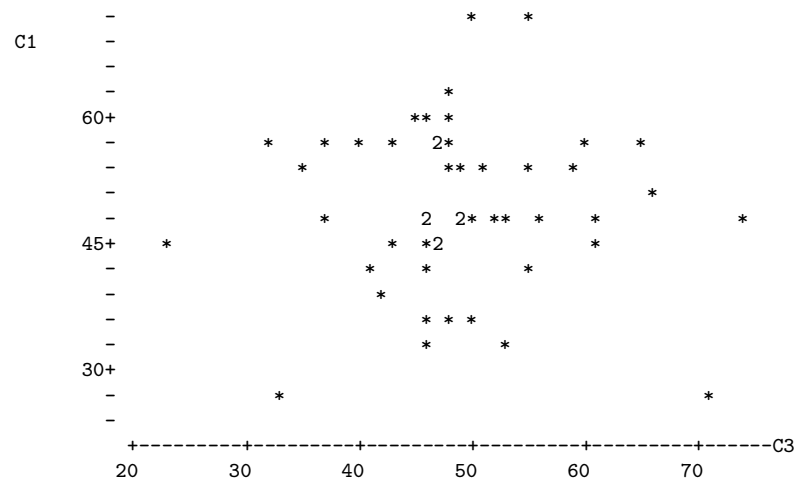
Often, the points in a scatterplot cluster around a straight line. The correlation coefficient (Pearson's r) expresses the extent to which the points cluster tightly around a straight line.

- $-1 \leq r \leq 1$
- $r = +1$ indicates a perfect positive linear relationship. All the points are exactly on a line with a positive slope.
- $r = -1$ indicates a perfect negative linear relationship. All the points are exactly on a line with a negative slope.
- $r = 0$ means no *linear* relationship (curve possible)
- r^2 represents explained variation, reduction in (squared) error of prediction. For example, the correlation between scores on the Scholastic Aptitude Test (SAT) and first-year grade point average (GPA) is around +0.50, so we say that SAT scores explain around 25% of the variation in first-year GPA.

The test of significance for Pearson's r assumes a bivariate normal distribution for the two variables; this means that the only possible relationship between them is linear. As usual, the assumption of independent observations is always important.

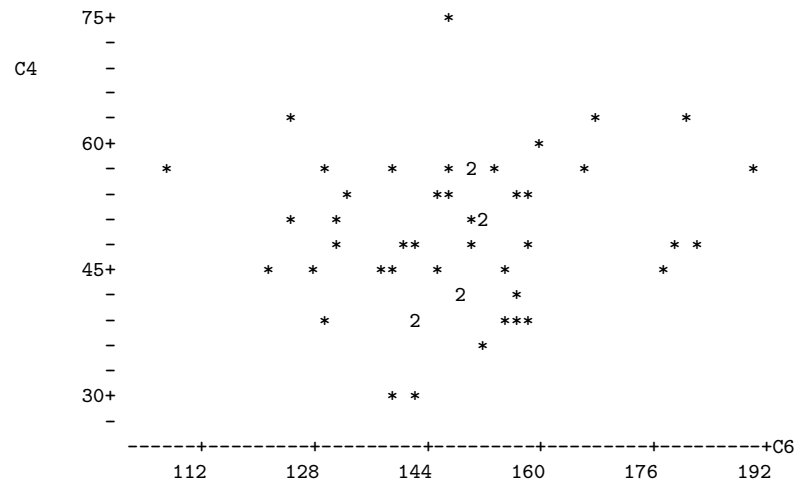
Here are some examples of scatterplots and the associated correlation coefficients.

```
MTB > plot c1 c3
```



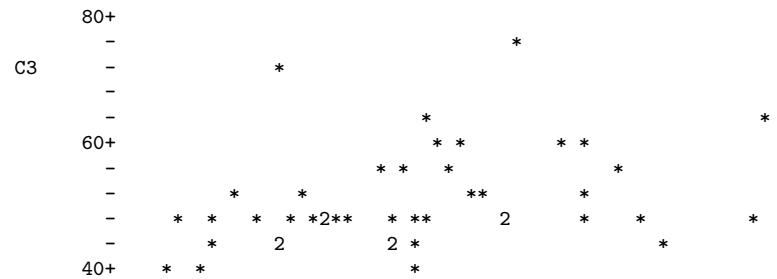

```
MTB > corr c1 c3
Correlation of C1 and C3 = 0.004
```

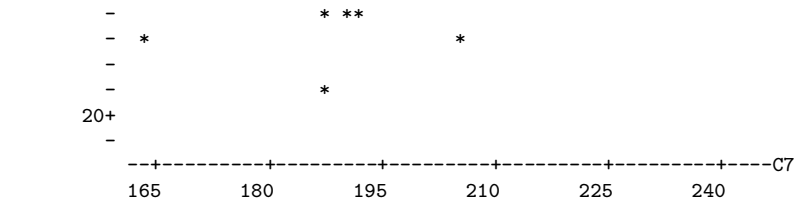
```
MTB > plot c4 c6
```




```
MTB > corr c4 c6
Correlation of C4 and C6 = 0.112
```

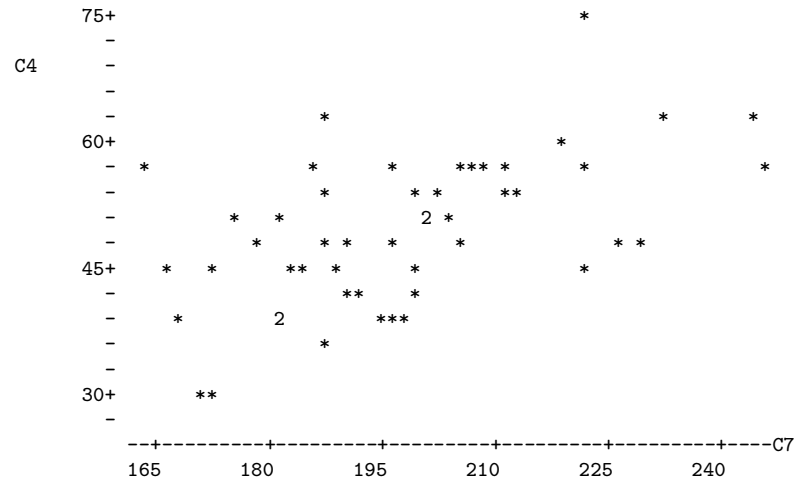
```
MTB > plot c3 c7
```





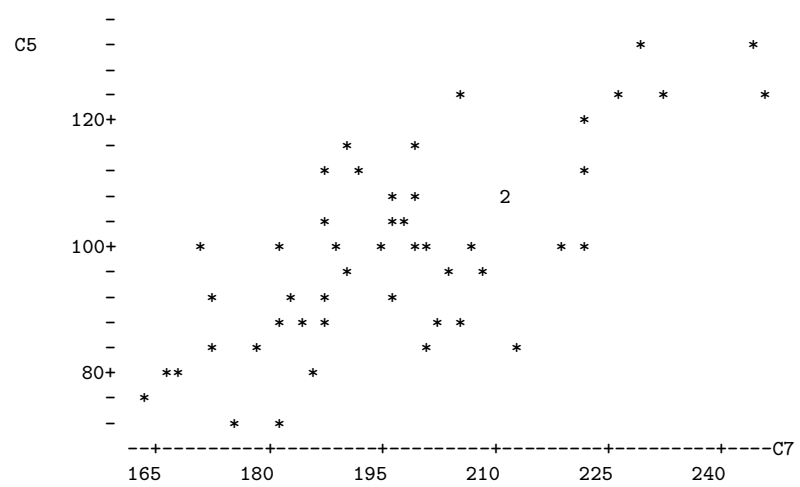

```
MTB > correlation between c3 and c7 please
Correlation of C3 and C7 = 0.368
```

```
MTB > plot c4 c7
```

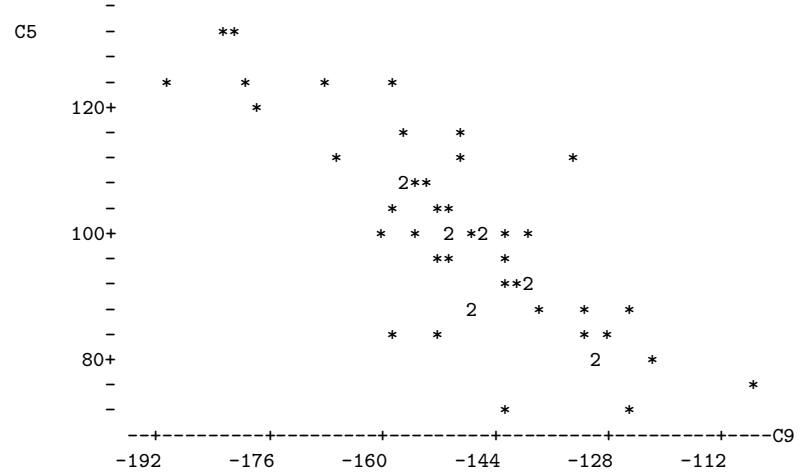



```
MTB > corr c4 c7
Correlation of C4 and C7 = 0.547
```

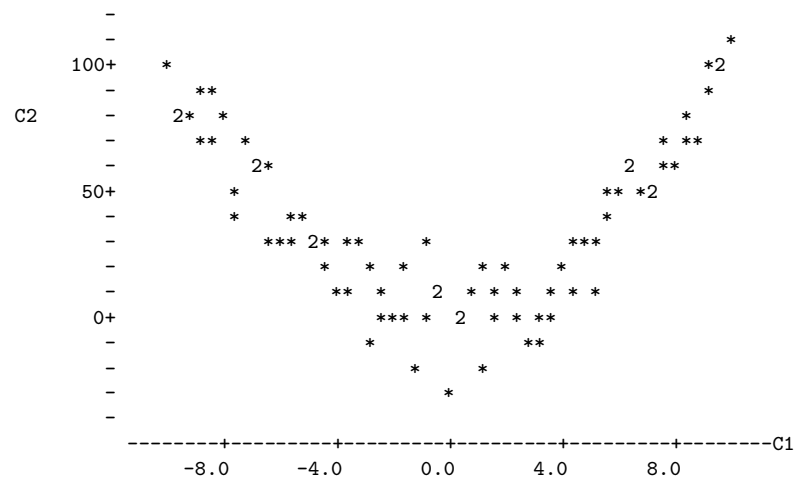
```
MTB > plot c5 c7
```



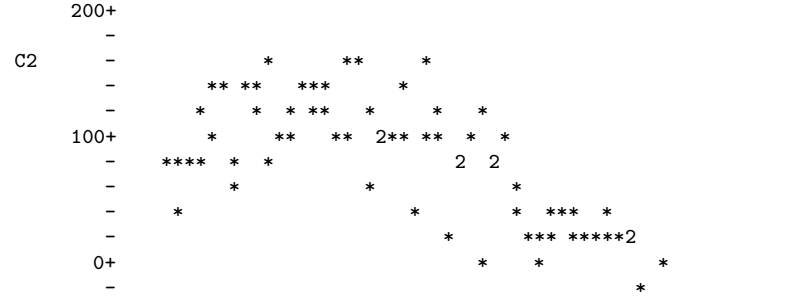
```
MTB > corr c5 c7
Correlation of C5 and C7 = 0.733


MTB > plot c5 c9
```

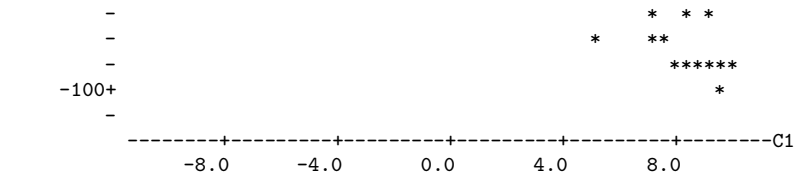



```
MTB > corr c5 c9
Correlation of C5 and C9 = -0.822




MTB > plot c2 c1
```




```
MTB > corr c1 c2
Correlation of C1 and C2 = 0.025
```






```
Correlation of C1 and C2 = -0.811
```

Simple Regression One independent variable, one dependent. In the usual examples both are quantitative (continuous). We fit a **least-squares** line to the cloud of points in a scatterplot. The least-squares line is the unique line that minimizes the sum of squared vertical distances between the line and the points in the scatterplot. That is, it minimizes the total (squared) error of prediction.

Denoting the slope of the least-squares line by b_1 and the intercept of the least-squares line by b_0,

$$b_1 = r\frac{s_y}{s_x} \text{ and } b_0 = \overline{Y} - b_1\overline{X}.$$

That is, the slope of the least squares has the same sign as the correlation coefficient, and equals zero if and only if the correlation coefficient is zero.

Usually, you want to test whether the slope is zero. This is the same as testing whether the correlation is zero, and mercifully yields the same p-value. Assumptions are independent observations (again) and that within levels of the IV, the DV has a normal distribution with the same variance (variance does not depend on value of the DV). Robustness properties are similar to those of the 2-sample t-test. The assumption of independent observations is always important.

Multiple Regression

Regression with several independent variables at once; we're fitting a (hyper) plane rather than a line. Multiple regression is very flexible; all the other techniques mentioned above (except the chi-squared test) are special cases of multiple regression. More details later.

Choosing an Elementary Technique

Make a table in lecture.

1.3 Experimental versus observational studies

Why might someone want to predict a dependent variable from an independent variable? There are two main reasons.

- There may be a practical reason for prediction. For example, a company might wish to predict who will buy a product, in order to maximize the productivity of its sales force. Or, an insurance company might wish to predict who will make a claim, or a university computer centre might wish to predict the length of time a type of hard drive will last before failing. In each of these cases, there will be some independent variables that are to be used for prediction, and although the people doing the study

may be curious and may have some ideas about how things might turn out and why, they don't really care why it works, as long as they can predict with some accuracy. Does variation in the IV *cause* variation in the DV? Who cares?

- This may be science (of some variety). The goal may be to understand how the world works — in particular, to understand the dependent variable. In this case, most likely we are implicitly or explicitly thinking of a causal relationship between the IV and DV. Think of attitude similarity and interpersonal attraction. . . .

Sample Question 1.3.1 *A study finds that high school students who have a computer at home get higher grades on average than students who do not. Does this mean that parents who can afford it should buy a computer to enhance their children's chances of academic success?*

Here is an answer that gets **zero** points. "Yes, with a computer the student can become computer literate, which is a necessity in our competitive and increasingly technological society. Also the student can use the computer to produce nice looking reports (neatness counts!), and obtain valuable information on the World Wide Web." **ZERO**.

The problem with this answer is that while it makes some fairly reasonable points, it is based on personal opinion, and fails to address the real question, which is "**Does this mean** . . . " Here is an answer that gets full marks.

Answer to Sample Question 1.3.1 *Not necessarily. While it is possible that some students are doing better academically and therefore getting into university because of their computers, it is also possible that their parents have enough money to buy them a computer, and also have enough money to pay for their education. It may be that an academically able student who is more likely to go to university will want a computer more, and therefore be more likely to get one somehow. Therefore, the study does not provide good evidence that a computer at home will enhance chances of academic success.*

Note that in this answer, the *focus is on whether the study provides good evidence* for the conclusion, not whether the conclusion is reasonable on other grounds. And the answer gives *specific alternative explanations* for the results as a way of criticizing the study. If you think about it, suggesting plausible alternative explanations is a very damaging thing to say about any empirical study, because you are pointing out that the investigators expended a huge amount of time and energy, but didn't establish anything conclusive. Also, suggesting alternative explanations is extremely valuable, because that is how research designs get improved and knowledge advances.

Now here are the general principles. If X and Y are measured at roughly the same time, X could be causing Y, Y could be causing X, or there might be some third variable (or collection of variables) that is causing both X and Y. Therefore we say that "Correlation does not necessarily imply causation." Here, by correlation we mean association (lack of independence) between variables. It is not limited to situations where you would compute a correlation coefficient.

A **confounding variable** is a variable not included as an independent variable, that might be related to both the independent variable and the dependent variable – and that might therefore create a seeming relationship between them where none actually exists,

or might even hide a relationship that is present. Some books also call this a "lurking variable." You are responsible for the vocabulary "confounding variable."

An **experimental study** is one in which cases are randomly assigned to the different values of an independent variable (or variables). An **observational study** is one in which the values of the independent variables are not randomly assigned, but merely observed.

Some studies are purely observational, some are purely experimental, and many are mixed. It's not really standard terminology, but in this course we will describe independent *variables* as experimental (i.e., randomly assigned, manipulated) or observed.

In an experimental study, there is no way the dependent variable could be causing the independent variable, because values of the IV are assigned by the experimenter. Also, it can be shown (using the Law of Large Numbers) that when units of observation are randomly assigned to values of an IV, all potential confounding variables are cancelled out as the sample size increases. This is very wonderful. You don't even have to know what they are!

Sample Question 1.3.2 *Is it possible for a continuous variable to be experimental, that is, randomly assigned?*

Answer to Sample Question 1.3.2 *Sure. In a drug study, let one of the independent variables consist of n equally spaced dosage levels spanning some range of interest, where n is the sample size. Randomly assign one participant to each dosage level.*

Sample Question 1.3.3 *Give an original example of a study with one quantitative observed independent variable and one categorical manipulated independent variable. Make the study multivariate, with one dependent variable consisting of unordered categories and two quantitative dependent variables. categorical*

Answer to Sample Question 1.3.3 *Stroke patients in a drug study are randomly assigned to either a standard blood pressure drug or one of three experimental blood pressure drugs. The categorical dependent variable is whether the patient is alive or not 5 years after the study begins. The quantitative dependent variables are systolic and diastolic blood pressure one week after beginning drug treatment.*

In practice, of course there would be a lot more variables; but it's still a good answer.

Because of possible confounding variables, only an experimental study can provide good evidence that an independent variable *causes* a dependent variable. Words like effect, affect, leads to etc. imply claims of causality and are only justified for experimental studies.

Sample Question 1.3.4 *Design a study that could provide good evidence of a causal relationship between having a computer at home and academic success.*

Answer to Sample Question 1.3.4 *High school students without computers enter a lottery. The winners (50% of the sample) get a computer and modem to use at home. The dependent variable is whether or not the student enters university.*

Sample Question 1.3.5 *Is there a problem with independent observations here? Can you fix it?*

Answer to Sample Question 1.3.5 *Oops. Yes. Students who win may be talking to each other, sharing software, etc.. Actually, the losers will be communicating too. Therefore their behaviour is non-independent and standard significance tests will be invalid. One solution is to hold the lottery in n separate schools, with one winner in each school. If the dependent variable were GPA, we could do a matched t-test comparing the performance of the winner to the average performance of the losers.*

Sample Question 1.3.6 *What if the DV is going to university or not?*

Answer to Sample Question 1.3.6 *We are getting into deep water here. Here is how I would do it. In each school, give a score of "1" to each student who goes to university, and a "0" to each student who does not. Again, compare the scores of the winners to the average scores of the losers in each school using a matched t-test. Note that the mean difference that is to be compared with zero here is the mean difference in probability of going to university, between students who get a computer to use and those who do not. While the differences for each school will not be normally distributed, the central limit theorem tells us that the mean difference will be approximately normal if there are more than about 20 schools, so the t-test is valid. In fact, the t-test is conservative, because the tails of the t distribution are heavier than those of the standard normal. This answer is actually beyond the scope of the present course.*

Artifacts and Compromises

Random assignment to experimental conditions will take care of confounding variables, but only if it is done right. It is amazingly easy for for confounding variables to sneak back into a true experimental study through defects in the procedure.

Placebo Effects

Experimenter Expectancy

Internal and external validity

Quasi-experimental designs

Chapter 2

First set of tools: SAS running under unix

The SAS language is the same regardless of what hardware you use or what operating system is running on the hardware. SAS programs are simple text files that can be transported from one machine to another with minimal difficulty. In this course, everything will be illustrated with SAS running under the unix operating system, but it's not a problem even if the next place you go only has PCs. It should take you about an hour to adjust to SAS-PC.

2.1 Unix

Unix is a line-oriented operating system. Well, there's X-windows (a graphical shell that runs on top of unix), but we won't bother with it. Basically, you type a command, press Enter, and unix does something for (or to) you. It may help to think of unix as DOS on steroids, if you remember DOS. The table below has all the unix commands you will need for this course. Throughout, *fname* stands for the name of a file.

A Minimal Set of unix Commands

exit Logs you off the system: ALWAYS log off before leaving!

passwd Lets you change your password. Recommended.

man *command name* Online help: explains *command name*, (like `man more`).

ls Lists names of the files in your directory.

more *fname* Displays fname on screen, one page at a time. Spacebar for next page, `q` to quit.

laser *fname* Prints hard copy on a laser printer. This is a local UTM command. The usual unix print command is `lpr` (for line printer).

draft *fname* Prints hard copy on a dot matrix printer. This is a local UTM command.

rm *fname* Removes *fname*, erasing it forever.

cp *fname1 fname2* Makes a copy of *fname1*. The new copy is named *fname2*.

mv *fname1 fname2* Moves (renames) *fname1*

pico *fname* Starts the `pico` text editor, editing *fname* (can be new file).

R Gets you into the `R` implementation of the S environment.

sas *fname* Executes `SAS` commands in the file ***fname*`.sas`**, yielding ***fname*`.log`** and (if no fatal errors) ***fname*`.lst`**.

ps Shows active processes

kill -9 # Kills process (job) number `#`. Sometimes you must do this when you can't log off because there are stopped jobs. Use `ps` to see the job numbers.

This really is a minimal set of commands. The unix operating system is extremely powerful, and has an enormous number of commands. You can't really see the power from the minimal set of commands above, but you can see the main drawback from the standpoint of the new user. Commands tend to be terse, consisting of just a few keystrokes. They make sense once you are familiar with them (like `ls` for listing the files in a directory, or `rm` for remove), but they are hard to guess. The `man` command (short for manual) gives very accurate information, but you have to know the name of the command before you can use `man` to find out about it.

Just for future reference, here are a few more commands that you may find useful, or otherwise appealing.

A Few More unix Commands

emacs *fname* Starts the `emacs` text editor, editing *fname* (can be new file). `Emacs` is much more powerful than `pico`.

mkdir *dirname* Makes a new sub-directory (like a folder) named *dirname*. You can have sub-directories within sub-directories; it's a good way to organize your work.

cp *fname dirname* Copies the file *fname* into the directory *dirname*.

cd *dirname* Short for Change Directory. Takes you to the sub-directory *dirname*.

cd .. Moves you up a directory level.

cd Moves you to your main directory from wherever you are.

ls > *fname* Sends the output of the `ls` command to the file *fname* instead of to the screen.

cat *fname* Lists the whole file on your screen, not one page at a time. It goes by very fast, but usually you can scroll back up to see the entire file, if it's not too long.

cat *fname1 fname2* > *fname3*

R –vanilla < *fname1* > *fname2*

grep ERROR cartoon1.log

alias chk "grep ERROR *.log ; grep WARN *.log"

cal

cal 1 3002

unset noclobber

rm -f *fname*

alias rm "rm -f"

rm -r *dirname*

rm *fname1 fname2*

Printing files at home This is a question that always comes up. Almost surely, the printer connected to your printer at home is not directly connected to the university network. If you want to do something like print your SAS output at home, you have to transfer the file on the unix machine to the hard drive of your home computer, and print it from there. You'll need to either use some kind of `ftp` (file transfer protocol) tool, or use the `more` or `cat` command to list the file on your screen, select it with your mouse, copy it, paste it to a word processing document, and print it from there.

2.2 Introduction to SAS

SAS stands for "Statistical Analysis System." Even though it runs on PCs and Macs as well as on bigger computers, it is truly the last of the great old mainframe statistical packages. The first beta release was in 1971, and the SAS Institute, Inc. was spun off from North Carolina State University in 1976, the year after Bill Gates dropped out of Harvard. This is a serious pedigree, and it has both advantages and disadvantages.

The advantages are that the number of statistical procedures SAS can do is truly staggering, and the most commonly used ones have been tested so many times by so many people that their correctness and numerical efficiency is beyond any question. For the purposes of this class, there are no bugs. The disadvantages of SAS are all related to the fact that it was *designed* to run in a batch-oriented mainframe environment. So, for example, the SAS Institute has tried hard to make SAS an "interactive" program, but has not really worked. It's as if someone painted an eighteen-wheel transport truck yellow, and called it a school bus. Yes, you can take the children to school in that thing, but would you want to?

2.2.1 The Four Main File Types

A typical SAS job will involve four main types of file.

- **The Raw Data File**: A file consisting of rows and columns of numbers; or maybe some of the columns have letters (character data) instead of numbers. The rows represent observations and the columns represent variables, as described at the beginning of Section 1.1. In the first example we will consider below, the raw data file is called `drp.dat`.

- **The Program File**: This is also sometimes called a "command file," because it's usually not much of a program. It consists of commands that the SAS software tries to follow. You create this file with a text editor like `pico` or `emacs`. The command file contains a reference to the raw data file (in the `infile` statement), so SAS knows where to find the data. In the first example we will consider below, the command file is called `reading.sas`. SAS expects program files to have the extension `.sas`, and you should always follow this convention.

- **The Log File**: This file is produced by every SAS run, whether it is successful of unsuccessful. It contains a listing of the command file, as well any error messages or warnings. The name of the log file is automatically generated by SAS; it combines the first part of the command file's name with the extension `.sas`. So for example, when SAS executes the commands in `reading.sas`, it writes a log file named `reading.log`.

- **The List File**: The list file contains the output of the statistical procedures requested by the command file. The list file has the extension `.lst` — so, for example, running SAS on the command file `reading.sas` will produce `reading.lst` as well as `reading.log`. A successful SAS run will almost always produce a list file. The absence of a list file indicates that there was at least one fatal error. The presence of a list file does not mean there were no errors; it just means that SAS was able to

do *some* of what you asked it to do. Even if there are errors, the list file will usually not contain any error messages; they will be in the log file.

2.2.2 Running SAS from the Command Line

There are several ways to run SAS. We will run SAS from the unix command line. In my view, this way is simplest and best.

If, by accident or on purpose, you type SAS without a filename, then SAS assumes you want to initiate an interactive session, and it tries to start the SAS Display Manager. If you are logged in through an ordinary telnet or ssh session, SAS terminates with an error: `ERROR: Cannot open X display. Check display name/server access authorization.` SAS assumes you are using the unix X-window graphical interface, so it will not work if your computer is emulating a (semi) dumb terminal. If you are in an X-window session, after a while several windows will open up. The only suggestion I have is this: Make sure the SAS Program Editor window is selected. From the File menu, choose Exit. Whew.

If you choose to ignore this advice and actually try to use the Display Manager, you are on your own. You will have my sympathy, but not my help. The joke about painting the transport truck yellow applies, and the joke is on you.

The following illustrates a simple SAS run from the command line. Initially, there are only two files in the (sub)directory — `reading.sas` (the program file) and `drp.dat` (the raw data file). The command `sas reading` produces two additional files — `reading.log` and `reading.lst`. In this and other examples, the unix prompt is `tuzo.erin` (the name of the unix machine used to produce the examples), followed by a `>` sign.

```
tuzo.erin > ls
drp.dat         reading.sas
tuzo.erin > sas reading
tuzo.erin > ls
drp.dat         reading.log     reading.lst     reading.sas
```

2.2.3 Structure of the Program File

A SAS program file is composed of units called *data steps* and *proc steps*. The typical SAS program has one data step and at least one proc step, though other structures are possible.

- Most SAS commands belong either in data step or in a proc step;they will generate errors if they are used in the wrong kind of step.
- Some statements, like the `title` and `options` commands, exist outside of the data and proc steps, but there are relatively few of these.

The Data Step The data step takes care of data acquisition and modification. It almost always includes a reference to the raw data file, telling SAS where to look for the data. It specifies variable names and labels, and provides instructions about how to read the data; for example, the data might be read from fixed column locations. Variables from the raw data file can be modified, and new variables can be created.

Each data step creates a **SAS data set**, a file consisting of the data (after modifications and additions), labels, and so on. Statistical procedures operate on SAS data sets, so you must create a SAS data set before you can start computing any statistics.

A SAS data set is written in a binary format that is very convenient for SAS to process, but is not readable by humans. In the old days, SAS data sets were always written to temporary scratch files on the computer's hard drive; these days, they may be maintained in RAM if they are small enough. In any case, the default is that a SAS data set disappears after the job has run. If the data step is executed again in a later run, the SAS data set is re-created.

Actually, it is possible to save a SAS data set on disk for later use. We won't do this much (there will be just one example), but it makes sense when the amount of processing in a data step is large relative to the speed of the computer. As an extreme example, one of my colleagues uses SAS to analyze data from Ontario hospital admissions; the data files have millions of cases. Typically, it takes around 20 hours of CPU time on a very strong unix machine just to read the data and create a SAS data set. The resulting file, hundreds of gigabytes in size, is saved to disk, and then it takes just a few minutes to carry out each analysis. You wouldn't want to try this on a PC.

To repeat, SAS data *steps* and SAS data *sets* sound similar, but they are distinct concepts. A SAS data *step* is part of a SAS program; it generates a SAS data *set*, which is a file – usually a temporary file.

SAS data sets are not always created by SAS data steps. Some statistical procedures can create SAS data sets, too. For example, `proc univariate` can take an ordinary SAS data set as input, and produce an output data set that has all the original variables, and also some of the variables converted to z-scores (by subtracting off the mean and dividing by the standard deviation). `Proc reg` (the main multiple regression procedure) can produce a SAS data set containing residuals for plotting and use in further analysis; there are many other examples.

The Proc Step "Proc" is short for procedure. Most procedures are statistical procedures; the main exception is `proc format`, which is used to provide labels for the values of categorical independent variables. The proc step is where you specify a statistical procedure that you want to carry out. A statistical procedures in the proc step will take a SAS data sets as input, and write the results (summary statistics, values of test statistics, p-values, and so on) to the list file. The typical SAS program includes one data step and several proc steps, because it is common to produce a variety of data displays, descriptive statistics and significance tests in a single run.

2.2.4 A First Example: `reading.sas`

Earlier, we ran SAS on the file `reading.sas`, producing `reading.log` and `reading.lst`. Now we will look at `reading.sas` in some detail. This program is very simple; it has just one data step and one proc step. More details will be given later, but it's based on a study in which one group of grade school students received a special reading programme, and a control group did not. After a couple of months, all students were given a reading test. We're just going to do an independent groups t-test, but first take a look at the raw data file. You'd do this with the unix `more` command.

Actually, it's so obvious that you should look at your data that nobody ever says it. But experienced data analysts always do it — or else they assume everything is okay and get a bitter lesson in something they already knew. It's so important that it gets the formal status of a **data analysis hint**.

Data Analysis Hint 1 *Always look at your raw data file. It the data file is big, do it anyway. At least page through it a screen at a time, looking for anything strange. Check the values of all the variables for a few cases. Do they make sense? If you have obtained the data file from somewhere, along with a description of what's in it, never believe that the description you have been given is completely accurate.*

Anyway, here is the file `drp.dat`, with the middle cut out to save space.

Treatment	24
Treatment	43
Treatment	58
⋮	⋮
Control	55
Control	28
Control	48
⋮	⋮

Now we can look at `reading.sas`.

```
/****************** reading.sas *********************
*  Simple SAS job to illustrate a two-sample t-test    *
*******************************************************/

options linesize=79 noovp formdlim='_';
title 'More & McCabe (1993) textbook t-test Example 7.8';

data reading;
     infile 'drp.dat';
     input group $ score;
     label group = 'Get Directed Reading Programme?'
           score = 'Degree of Reading Power Test Score';
proc ttest;
     class group;
     var score;
```

Here are some detailed comments about `reading.sas`.

- The first three lines are a comment. Anything between a `/*` and `*/` is a comment, and will be listed on the log file but otherwise ignored by SAS. Comments can appear anywhere in a program. You are not required to use comments, but it's a good idea.

The most common error associated with comments is to forget to end them with `*/`. In the case of `reading.sas`, leaving off the `*/` (or typing
by mistake) would cause the whole program to be treated as a comment. It would generate no errors, and no output — because as far as SAS would be concerned, you never requested any. A longer program would eventually exceed the default length of a comment (it's some large number of characters) and SAS would end the "comment" for you. At exactly that point (probably in the middle of a command) SAS would begin parsing the program. Almost certainly, the first thing it examined would be a fragment of a legal command, and this would cause an error. The log file would say that the command caused an error, and not much else. It would be *very* confusing, because probably the command would be okay, and there would be no indication that SAS was only looking at part of it.

- The next two lines (the `options` statement and the `title` statement) exist outside the proc step and the data step. This is fairly rare.

- All SAS statements end with a semi-colon (`;`). SAS statements can extend for several physical lines in the program file (for example, see the `label` statement). Spacing, indentation, breaking up s statement into several lines of text – these are all for the convenience of the human reader, and are not part of the SAS syntax.

- The most common error in SAS programming is to forget the semi-colon. When this happens, SAS tries to interpret the following statement as part of the one you tried to end. This often causes not one error, but a cascading sequence of errors. The rule is, *if you have an error and you do not immediately understand what it is, look for a missing semi-colon.* It will probably be *before* the portion of the program that (according to SAS) caused the first error.

- Cascading errors are not caused just by the dreaded missing semi-colon. They are common in SAS; for example, a runaway comment statement can easily cause a chain reaction of errors (if the program is long enough for it to cause any error messages at all). *If you have a lot of errors in your log file, fix the first one and don't waste time trying to figure out the others.* Some or all of them may well disappear.

- `options linesize=79 noovp formdlim='_';`

  These options are highly recommended. The `linesize=79` option is so highly recommended it's almost obligatory. It causes SAS to write the output 79 columns across, so it can be read on an ordinary terminal screen that's 80 characters across. You specify an output width of 79 characters rather than 80, because SAS uses one column for printer control characters, like page ejects (form feeds).

  If you do not specify `options linesize=79;`, SAS will use its default of 132 characters across, the width of sheet of paper from an obsolete line printer you probably have never seen. Why would the SAS Institute hang on to this default, when changing it to match ordinary letter paper would be so easy? It probably tells you something about the computing environments of some of SAS's large corporate clients.

- The `noovp` option makes the log files more readable if you have errors. When SAS finds an error in your program, it tries to *underline* the word that caused the error. It does this by going back and *overprinting* the offending word with a series of "underscores" (_ characters). On many printers this works, but when you try to look at the log file on a terminal screen (one that is *not* controlled by the SAS Display Manager), what often appears is a mess. The `noovp` option specifies `no overprinting`. It causes the "underlining" to appear on a separate line under the program line with the error. If you're running SAS from the unix command line and looking at your log files with the `more` command (or the `less` command or the `cat` commmand), you will probably find the `noovp` option to be helpful.

- The `formdlim='_'` option specifies a "form delimiter" to replace most form feeds (new physical pages) in the list file. This can save a lot of paper (and page printing charges). You can use any string you want for a form delimiter. The underscore (the one specified here) causes a solid line to be printed instead of going to a new sheet of paper.

- `title` This is optional, but recommended. The material between the single quotes will appear at the top of each page. This can be a lifesaver when you are searching through a stack of old printouts for something you did a year or two ago.

- `data reading;` This begins the data step, specifying the name of the SAS data set that is being created.

- `infile` Specifies the name of the raw data file. The file name, enclosed in single quotes, can be the full unix path to the file, like `/dos/brunner/public/senic.raw`. If you just give the name of the raw data file, as in this example, SAS assumes that the file is in the same directory as the command file.

- `input` Gives the names of the variables.

  - A character variable (the values of `group` are "Treatment' and "Control") must be followed by a dollar sign.
  - Variable names must be eight characters or less, and should begin with a letter. They will be used to request statistical procedures in the `proc` step. They should be meaningful (related to what the variable *is*), and easy to remember.
  - This is almost the simplest form of the `input` statement. It can be very powerful; for example, you can read data from different locations and in different orders, depending on the value of a variable you've just read, and so on. It can get complicated, but if the data file has a simple structure, the input statement can be simple too.

- `label` Provide descriptive labels for the variables; these will be used to label the output, usually in very nice way. Labels can be quite useful, especially when you're trying to recover what you did a while ago. Notice how this statement extends over two physical lines.

- `proc ttest;` Now the proc step begins. This program has only one data step and one proc step. We are requesting a two-sample t-test.

- `class` Specifies the independent variable.
- `var` Specifies the dependent variable(s). You can give a list of dependent variables. A separate univariate test (actually, as you will see, *collection* of tests is performed for each dependent variable.

reading.log Log files are not very interesting when everything is okay, but here is an example anyway. Notice that in addition to a variety of technical information (where the files are, how long each step took, and so on), it contains a listing of the SAS program — in this case, `reading.sas`. If there were syntax errors in the program, this is where the error messages would appear.

```
tuzo.erin > cat reading.log
1                                                       The SAS System                              11:08 Friday, January 2,

NOTE: Copyright (c) 1989-1996 by SAS Institute Inc., Cary, NC, USA.
NOTE: SAS (r) Proprietary Software Release 6.12  TS020
      Licensed to UNIVERSITY OF TORONTO/COMPUTING & COMMUNICATIONS, Site 0008987001.



This message is contained in the SAS news file, and is presented upon
initialization.  Edit the files "news" in the "misc/base" directory to
display site-specific news and information in the program log.
The command line option "-nonews" will prevent this display.



NOTE: AUTOEXEC processing beginning; file is /local/sas612/autoexec.sas.

NOTE: SAS initialization used:
      real time           0.780 seconds
      cpu time            0.152 seconds


NOTE: AUTOEXEC processing completed.

1          /****************** reading.sas *********************
2          *  Simple SAS job to illustrate a two-sample t-test    *
3          ******************************************************/
4
5          options linesize=79 noovp formdlim='_';
6          title 'More & McCabe (1993) textbook t-test Example 7.8';
7          data reading;
8               infile 'drp.dat';
9               input group $ score;
10              label group = 'Get Directed Reading Programme?'
11                    score = 'Degree of Reading Power Test Score';

NOTE: The infile 'drp.dat' is:
      File Name=/res/jbrunner/442s04/notesSAS/drp.dat,
      Owner Name=jbrunner,Group Name=research,
      Access Permission=rw-------,
      File Size (bytes)=660

NOTE: 44 records were read from the infile 'drp.dat'.
      The minimum record length was 14.
      The maximum record length was 14.
NOTE: The data set WORK.READING has 44 observations and 2 variables.
NOTE: DATA statement used:
      real time           0.190 seconds
      cpu time            0.051 seconds


12         proc ttest;
```

```
13              class group;
14              var score;
NOTE: The PROCEDURE TTEST printed page 1.
NOTE: PROCEDURE TTEST used:
      real time           0.030 seconds
      cpu time            0.009 seconds


2                               The SAS System    11:08 Friday, January 2, 2004


NOTE: The SAS System used:
      real time           1.120 seconds
      cpu time            0.233 seconds

NOTE: SAS Institute Inc., SAS Campus Drive, Cary, NC USA 27513-2414
```

reading.lst Here is the list file. Notice that the title specified in the `title` statement appears at the top, along with the time and date the program was executed. Then we get means and standard deviations, and several statistical tests — including the one we wanted. We get other stuff too, whether we want it or not. This is typical of SAS, and most other mainstream statistical packages as well. The default output from any given statistical procedures will contain more information than you wanted, and probably some stuff you don't understand at all. There are usually numerous options that can add *more* information, but almost never options to reduce the default output. So, you just learn what to ignore. It is helpful, but not essential,to have at least a superficial understanding of everything in the default output from procedures you use a lot.

```
---------------------------------------------------------------------------

               More & McCabe (1993) textbook t-test Example 7.8              1
                                          11:08 Friday, January 2, 2004

                                TTEST PROCEDURE

Variable: SCORE        Degree of Reading Power Test Score

GROUP          N                 Mean              Std Dev            Std Error
---------------------------------------------------------------------------
Control       23          41.52173913          17.14873323          3.57575806
Treatmen      21          51.47619048          11.00735685          2.40200219

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal    -2.3109     37.9      0.0264
Equal      -2.2666     42.0      0.0286

For H0: Variances are equal, F' = 2.43    DF = (22,20)    Prob>F' = 0.0507
```

Now here are some comments about `reading.lst`.

- `Variable: SCORE` This tells you what the dependent variable is – particularly useful if you have more than one. Notice the nice use of the variable label that was supplied in the `label` statement.
- `GROUP` The independent variable. Underneath are the values of the independent variable. We also have the sample size n for each group, and the group mean, standard deviation, and also the standard error or the mean ($\frac{s}{\sqrt{n}}$, the estimated standard deviation of the sampling distribution of the sample mean).

- Well actually, if you look carefully, you see that we do *not* quite get the values of the independent variable under `GROUP`. The values of the (alphanumeric, or character-valued) variable `group` are `Control` and `Treatment`, but the printout says "`Treatmen`." This is not a printing error; it is a subtle error in the reading of the data. The default length of an alphanumeric data value is 8 characters, but "Treatment" has 9 characters. So SAS just read the first eight. No error message was generated and no harm was done in this case, but in other circumstances this error can turn a data file into a giant pile of trash, without warning. Later we will see how to override the default and read longer strings if necessary.

- Next we get a table whose first column is entitled "Variances." This gives t statistics for testing equality of means, which was what we are interested in. The traditional t-test assumes equal variances, and it is given in the column entitled "Equal."

  - The value of the test statistic is `-2.2666`.
  - The degrees of freedom $n_1 + n_2 - 2$ is given in the `DF` column.
  - The column `Prob>|T|` gives the two-tailed (two-sided) p-value. It is less than the traditional value of 0.05, so the results are statistically significant.

Sample Question 2.2.1 *What do we conclude from this study? Say something about reading, using non-technical language.*

Answer to Sample Question 2.2.1 *Students who received the Directed Reading Program got higher average reading scores than students in the control condition.*

It's worth emphasizing here that the main objective of doing a statistical analysis is to draw conclusions about the data — or to refrain from drawing such conclusions, for good reasons. The question "What do we conclude from this study?" will always be asked. The right answer will always be either "Nothing; the results were not statistically significant," or else it will be something about reading, or fish, or potatoes, or AIDS, or whatever is being studied. Many students, even when they have been warned, respond with a barrage of statistical terminology. They go on and on about the null hypothesis and Type I error, and usually say nothing that would tell a reasonable person what actually happened in the study. In the working world, a memo filled with such garbage could get you fired. Here, it will get you a zero for the question, even if the technical details you give are correct.

Remember, the purpose of writing up a statistical analysis is not to sound impressive and technical, but to impart information. To say things in a simple way is a virtue. It shows you understand what is going on. Now back to the printout.

- The row entitled "Unequal" gives a sort of t-test that does not assume equal variances. Well, it's not really a t-test, because the test statistic does not really have a t distribution, even when the data are exactly normal. But, the (very unpleasant) distribution of the test statistic is well approximated by a t distribution with the right degrees of freedom — not $n_1 + n_2 - 2$, but something messy that depends on the data. See the odd fractional degrees of freedom? See [2] for details. In any case, it does not matter much in this case, because the p-value is almost the same as the

p-value from the traditional test. They lead to the same conclusions, and there is no problem. What should you do when they disagree? I'd go with the test that makes fewer assumptions.

- Next we see `For H0: Variances are equal` and an F-test. This is the traditional test for whether the variances of two groups are equal, and it's *almost* significant. This test is provided so people can test for differences between variances; if it is significantly different they can use the unequal variance t-test, and otherwise they can use the traditional test. This seems reasonable, except for the following.

  Both the two-sample t-test and the F-test for equality of variances assume that the data are normally distributed. However, the normality assumption does not matter much for the t-test when the sample sizes are large, while for the variance test it matters a *lot*, regardless of how much data you have. When the data are non-normal, the test for variances will be significant more than 5% of the time even when the population variances are equal. If you have equal population variances and a large sample of non-normal data, the F-test for variances could easily be significant, leading you to worry unnecessarily about the validity of the t-test.

2.2.5 Background of the First Example

We don't do statistical analysis in a vacuum. Before proceeding with more computing details, let's find out more about the reading data. This first example is from an introductory text. It's Example 7.8 (p. 534) in More and McCabe's excellent *Introduction to the practice of statistics* [2]. We are interested in analyzing *real* data, not in doing textbook exercises. But we will not turn up our noses just yet, because

Data Analysis Hint 2 *When learning how to carry out a procedure using unfamiliar statistical software, always do a textbook example first, and compare the output to the material in the text. Regardless of what the manual might say, never assume you know what the software is doing until you see an example.*

More and McCabe do a great job of explaining the t-test with unequal variances, something SAS produces (along with usual t-test that assumes equal variances) without being asked when you request a t-test. Besides, the data actually come from someone's Ph.D. thesis, so there is an element of realism. Here is Moore and McCabe's description of the study.

> An educator believes that new directed reading activities in the classroom will help elementary school pupils improve some aspects of their reading ability. She arranges for a third grade class of 21 students to take part in these activities. A control classroom of 23 third graders follows the same curriculum without the activities. At the end of 8 weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the program is designed to improve.

Sample Question 2.2.2 *What's wrong with this study?*

Answer to Sample Question 2.2.2 *The independent variable was manipulated by the experimenter, but it is not an experimental study. Even if classrooms were assigned randomly to conditions (it is impossible to tell whether they were, from this brief description), a large number of unobserved variables are potentially confounded with treatment. The teacher in the classroom that received the treatment might be better than the teacher in the control classroom, or possibly there was a particularly aggressive bully in the control classroom, or maybe a mini-epidemic of some childhood disease hit the control classroom —vdots. The list goes on. The point here is that there are many ways in which the classroom experiences of children in the treatment group differ systematically from the experiences of children in the control group.*

Sample Question 2.2.3 *How could the problem be fixed?*

Answer to Sample Question 2.2.3 *Assign classrooms at random to treatments. The unit of analysis should be the classroom, not the individual student.*

2.2.6 SAS Example Two: The statclass data

These data come from a statistics class taught many years ago. Students took eight quizzes, turned in nine computer assignments, and also took a midterm and final exam. The data file also includes gender and ethnic background; these last two variables are just guesses by the professor, and there is no way to tell how accurate they were. The data file looks like this. There are 21 columns and 62 rows of data; columns not aligned.

```
tuzo.erin >  more statclass.dat
1  2  9  1  7  8  4  3  5  2  6  10  10  10  5  0  0  0  0  55  43
0  2  10  10  5  9  10  8  6  8  10  10  8  9  9  9  9  10  10  66  79
1  2  10  10  5  10  10  10  9  8  10  10  10  10  10  10  9  10  10  94  67
1  2  10  10  8  9  10  7  10  9  10  10  10  9  10  10  9  10  10  81  65
0  1  10  1  0  0  8  6  5  2  10  9  0  0  10  6  0  5  0  54  29
```

⋮

Here is the SAS program.

```
tuzo.erin >  cat statmarks.sas
options linesize=79 pagesize=35;
title 'Grades from STA3000 at Roosevelt University:  Fall, 1957';
title2 'Illustrate Elementary Tests';

proc format; /* Used to label values of the categorical variables */
     value sexfmt    0 = 'Male'   1 = 'Female';
     value ethfmt    1 = 'Chinese'
                     2 = 'European'
                     3 = 'Other' ;
data grades;
     infile 'statclass.dat';
     input sex ethnic quiz1-quiz8 comp1-comp9 midterm final;
     /* Drop lowest score for quiz & computer  */
     quizave = ( sum(of quiz1-quiz8) - min(of quiz1-quiz8) ) / 7;
     compave = ( sum(of comp1-comp9) - min(of comp1-comp9) ) / 8;
     label ethnic = 'Apparent ethnic background (ancestry)'
           quizave = 'Quiz Average (drop lowest)'
           compave = 'Computer Average (drop lowest)';
     mark = .3*quizave*10 + .1*compave*10 + .3*midterm + .3*final;
     label mark = 'Final Mark';
     diff = quiz8-quiz1; /* To illustrate matched t-test */
     label diff = 'Quiz 8 minus Quiz 1';
     format sex sexfmt.;       /* Associates sex & ethnic    */
     format ethnic ethfmt.;    /* with formats defined above */

proc freq;
     tables sex ethnic;
proc means n mean std;
     var quiz1 -- mark;        /*  single dash only works with numbered
                                  lists, like quiz1-quiz8    */
proc ttest;
     title 'Independent t-test';
     class sex;
     var mark;
proc means n mean std t;
     title 'Matched t-test: Quiz 1 versus 8';
     var quiz1 quiz8 diff;
proc glm;
     title 'One-way anova';
     class ethnic;
     model mark = ethnic;
     means ethnic / Tukey Bon Scheffe;
proc freq;
     title 'Chi-squared Test of Independence';
     tables sex*ethnic / chisq;
proc freq; /*  Added after seeing warning from chisq test above */
```

```
     title 'Chi-squared Test of Independence: Version 2';
     tables sex*ethnic / norow nopercent chisq expected;
proc corr;
     title 'Correlation Matrix';
     var final midterm quizave compave;
proc plot;
     title 'Scatterplot';
     plot final*midterm; /* Really should do all combinations */
proc reg;
     title 'Simple regression';
     model final=midterm;

/* Predict final exam score from midterm, quiz & computer */
proc reg simple;
     title 'Multiple Regression';
     model final = midterm quizave compave / ss1;
     smalstuf: test quizave = 0, compave = 0;
```

Noteworthy features of this program include

- `options linesize=79 pagesize=35;` Good for $8\frac{1}{2}$ by 11 paper.
- `title2` Subtitle
- `proc format`
- `quiz1-quiz8`
- Creating new variables with assignment statements
- `sum(of quiz1-quiz8)`
- `diff = quiz8-quiz1`
- `format sex sexfmt.;`
- `quiz1 -- mark`
- Title inside a procedure labels just that procedure
- `proc freq` For frequency distributions
- `proc means` To get means and standard deviations
- `proc ttest` We've seen
- `proc means n mean std t` A matched t-test is just a single-variable t-test carried out on differences, testing whether the mean difference is equal to zero.
- `proc glm`
  - `class` Tells SAS that ethnic is categorical.

  - `model` Dependent variable(s) = independent variable(s)
  - `means ethnic / Tukey Bon Scheffe`

- `chisq` option on `proc freq`
- `chisq` option on `proc freq`
- `tables sex*ethnic / norow nopercent chisq expected;` In version 2 of `proc freq`
- `proc corr`
- `proc plot; plot final*midterm;` Scatterplot: First variable named goes on the y axis.
- `proc reg`: `model` Dependent variable(s) = independent variable(s) again
- `simple` option on `proc reg` gives simple descriptive statistics. This last procedure is an example of multiple regression, and we will return to it later once we have more background.

statmarks.lst

```
          Grades from STA3000 at Roosevelt University:  Fall, 1957          1
                        Illustrate Elementary Tests
                                              10:20 Friday, January 4, 2002

                                          Cumulative  Cumulative
             SEX   Frequency   Percent   Frequency    Percent
            ---------------------------------------------------
            Male           39      62.9          39       62.9
            Female         23      37.1          62      100.0

                    Apparent ethnic background (ancestry)

                                          Cumulative  Cumulative
              ETHNIC   Frequency   Percent   Frequency    Percent
            ---------------------------------------------------
            Chinese          41      66.1          41       66.1
            European         15      24.2          56       90.3
            Other             6       9.7          62      100.0

^L          Grades from STA3000 at Roosevelt University:  Fall, 1957
2
                        Illustrate Elementary Tests
                                              10:20 Friday, January 4, 2002

   Variable  Label                              N          Mean       Std Dev
   ------------------------------------------------------------------------
   QUIZ1                                       62     9.0967742     2.2739413
   QUIZ2                                       62     5.8870968     3.2294995
   QUIZ3                                       62     6.0483871     2.3707744
   QUIZ4                                       62     7.7258065     2.1590022
   QUIZ5                                       62     9.0645161     1.4471109
   QUIZ6                                       62     7.1612903     1.9264641
   QUIZ7                                       62     5.7903226     2.1204477
   QUIZ8                                       62     6.3064516     2.3787909
   COMP1                                       62     9.1451613     1.1430011
   COMP2                                       62     8.8225806     1.7604414
   COMP3                                       62     8.3387097     2.5020880
   COMP4                                       62     7.8548387     3.2180168
   COMP5                                       62     9.4354839     1.7237109
   COMP6                                       62     7.8548387     2.4350364
```

```
   COMP7                                 62     6.6451613     2.7526248
   COMP8                                 62     8.8225806     1.6745363
   COMP9                                 62     8.2419355     3.7050497
   MIDTERM                               62    70.1935484    13.6235557
   FINAL                                 62    49.4677419    17.5141327
   QUIZAVE   Quiz Average (drop lowest)      62     7.6751152     1.1266917
   COMPAVE   Computer Average (drop lowest)  62     8.8346774     1.1204997
   MARK      Final Mark                  62    67.7584101    11.0235746
   ------------------------------------------------------------------------
^L                             Independent t-test
3
                                                   10:20 Friday, January 4, 2002

                               TTEST PROCEDURE

Variable: MARK         Final Mark

   SEX       N         Mean      Std Dev    Std Error      Minimum      Maximum
--------------------------------------------------------------------------------
  Male      39  67.62097070  10.11112521   1.61907581  43.61428571  89.93214286
Female      23  67.99145963  12.65945704   2.63967927  48.48214286  95.45714286

Variances        T       DF    Prob>|T|
---------------------------------------
Unequal    -0.1196     38.5      0.9054
Equal      -0.1268     60.0      0.8995

For H0: Variances are equal, F' = 1.57    DF = (22,38)    Prob>F' = 0.2190
^L                        Matched t-test: Quiz 1 versus 8
4
                                                   10:20 Friday, January 4, 2002

  Variable  Label                 N          Mean       Std Dev            T
  ------------------------------------------------------------------------
  QUIZ1                          62     9.0967742     2.2739413    31.4995252
  QUIZ8                          62     6.3064516     2.3787909    20.8749114
  DIFF      Quiz 8 minus Quiz 1  62    -2.7903226     3.1578011    -6.9576965
  ------------------------------------------------------------------------
^L                                One-way anova
5
                                                   10:20 Friday, January 4, 2002

                        General Linear Models Procedure
                            Class Level Information

                   Class    Levels    Values

                   ETHNIC        3    Chinese European Other

                    Number of observations in data set = 62

^L                                One-way anova
6
                                                   10:20 Friday, January 4, 2002

                        General Linear Models Procedure

Dependent Variable: MARK   Final Mark
                                     Sum of            Mean
Source                  DF          Squares          Square  F Value    Pr > F

Model                    2     1478.9595320     739.4797660     7.35    0.0014

Error                   59     5933.7115164     100.5713816

Corrected Total         61     7412.6710484

                  R-Square             C.V.        Root MSE         MARK Mean

                  0.199518         14.80042       10.028528         67.758410
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
ETHNIC	2	1478.9595320	739.4797660	7.35	0.0014

Source	DF	Type III SS	Mean Square	F Value	Pr > F
ETHNIC	2	1478.9595320	739.4797660	7.35	0.0014

General Linear Models Procedure

Tukey's Studentized Range (HSD) Test for variable: MARK

NOTE: This test controls the type I experimentwise error rate.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of Studentized Range= 3.400

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.108	6.539	18.185	
European - Chinese	4.252	11.528	18.803	***
Other - European	-18.185	-6.539	5.108	
Other - Chinese	-5.550	4.989	15.528	
Chinese - European	-18.803	-11.528	-4.252	***
Chinese - Other	-15.528	-4.989	5.550	

General Linear Models Procedure

Bonferroni (Dunn) T tests for variable: MARK

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of T= 2.46415

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.398	6.539	18.476	
European - Chinese	4.071	11.528	18.985	***
Other - European	-18.476	-6.539	5.398	
Other - Chinese	-5.813	4.989	15.790	
Chinese - European	-18.985	-11.528	-4.071	***
Chinese - Other	-15.790	-4.989	5.813	

General Linear Models Procedure

Scheffe's test for variable: MARK

NOTE: This test controls the type I experimentwise error rate but generally has a higher type II error rate than Tukey's for all pairwise comparisons.

Alpha= 0.05 Confidence= 0.95 df= 59 MSE= 100.5714
Critical Value of F= 3.15312

Comparisons significant at the 0.05 level are indicated by '***'.

ETHNIC Comparison	Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
European - Other	-5.626	6.539	18.704	
European - Chinese	3.928	11.528	19.127	***
Other - European	-18.704	-6.539	5.626	
Other - Chinese	-6.019	4.989	15.997	
Chinese - European	-19.127	-11.528	-3.928	***
Chinese - Other	-15.997	-4.989	6.019	

TABLE OF SEX BY ETHNIC

SEX ETHNIC(Apparent ethnic background (ancestry))

Frequency| Expected | Col Pct	Chinese	European	Other	Total
Male	27	7	5	39
	25.79	9.4355	3.7742	
	65.85	46.67	83.33	
Female	14	8	1	23
	15.21	5.5645	2.2258	
	34.15	53.33	16.67	
Total	41	15	6	62

STATISTICS FOR TABLE OF SEX BY ETHNIC

Statistic	DF	Value	Prob
Chi-Square	2	2.921	0.232
Likelihood Ratio Chi-Square	2	2.996	0.224
Mantel-Haenszel Chi-Square	1	0.000	0.995
Phi Coefficient		0.217	
Contingency Coefficient		0.212	
Cramer's V		0.217	

Sample Size = 62
WARNING: 33% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

Correlation Analysis

4 'VAR' Variables: FINAL MIDTERM QUIZAVE COMPAVE

Simple Statistics

Variable	N	Mean	Std Dev	Sum
FINAL	62	49.467742	17.514133	3067.000000
MIDTERM	62	70.193548	13.623556	4352.000000
QUIZAVE	62	7.675115	1.126692	475.857143
COMPAVE	62	8.834677	1.120500	547.750000

Simple Statistics

Variable	Minimum	Maximum	Label
FINAL	15.000000	89.000000	
MIDTERM	44.000000	103.000000	
QUIZAVE	4.571429	9.714286	Quiz Average (drop lowest)

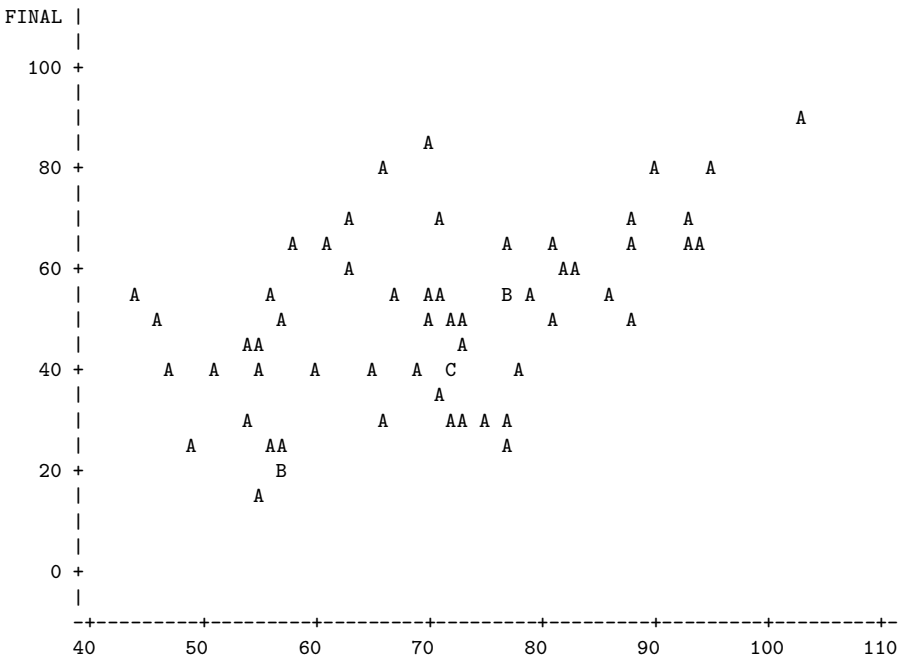
COMPAVE	5.000000	10.000000	Computer Average (drop lowest)

Correlation Analysis

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 62

	FINAL	MIDTERM	QUIZAVE	COMPAVE
FINAL	1.00000	0.51078	0.47127	0.14434
	0.0	0.0001	0.0001	0.2630
MIDTERM	0.51078	1.00000	0.59294	0.41277
	0.0001	0.0	0.0001	0.0009
QUIZAVE	0.47127	0.59294	1.00000	0.52649
Quiz Average (drop lowest)	0.0001	0.0001	0.0	0.0001
COMPAVE	0.14434	0.41277	0.52649	1.00000
Computer Average (drop lowest)	0.2630	0.0009	0.0001	0.0

Plot of FINAL*MIDTERM. Legend: A = 1 obs, B = 2 obs, etc.



Model: MODEL1
Dependent Variable: FINAL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	4881.79529	4881.79529	21.180	0.0001
Error	60	13829.64019	230.49400		
C Total	61	18711.43548			

Root MSE	15.18203	R-square	0.2609
Dep Mean	49.46774	Adj R-sq	0.2486
C.V.	30.69077		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	3.375101	10.19938324	0.331	0.7419
MIDTERM	1	0.656651	0.14268372	4.602	0.0001

Multiple regression output was deleted.

2.2.7 SAS Example Two: The SENIC data

These data are from a disk that comes with Neter et al's [3] *Applied linear statistical models*. The acronym SENIC stands for "Study of Nosocomial Infection Control." "Nosocomial" means acquired in hospital. Sometimes, patients go to hospital with a broken leg or something, and catch a severe respiratory infection, presumably from other patients. The observations here are hospitals, and the dependent variable is `infrisk`, the probability of catching an infection while in hospital (multiplied by 100). The other variables are explained fairly well by the `labels` statement.

First we will look at the file `senic0.sas`. This is a very basic program that just reads the data and does frequency distributions of everything (even identification number; you don't want to print this!). The idea is that you start out this way, checking for data errors, and then gradually build up the program, adding labels, printing formats and new variables a little bit at a time. This makes it easier to catch your errors.

```
/* senic0.sas */
options linesize = 79;
data  simple;
      infile 'senic.dat';
      input   id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;
proc freq;
      tables _all_;
```

Now suppose we discovered that the file has some weird missing value codes. The next version of the program might look like this.

```
/* senic0.1.sas */
options linesize = 79;
data  simple;
      infile 'senic.dat';
      input   id stay age infrisk culratio xratio nbeds medschl
              region census nurses service;

      /*** sas doesn't like numeric missing value codes.  a period  .  is
                      best for missing. however ....  ***/

      if stay eq 9999 then stay = .  ;
      if age eq 9999 then age = .  ;
      if xratio eq 9999 then xratio = .  ;
      if culratio eq 9999 then culratio = .  ;
      if infrisk = 999 then infrisk  = .  ;
      if nbeds = 9 then nbeds  = .  ;
      if medschl = 9 then medschl  = .  ;
      if region = 9 then region  = .  ;
      if census = 9 then census  = .  ;
      if service = 9 then service  = .  ;
```

```
      if nurses eq (0 or .999) then nurses = .  ;

proc freq;
      tables _all_;
```

The process continues. On the way, we switch to a version of the data file that has the data lined up in fixed columns, with blanks for missing values (a common situation). We wind up with a program called `senicread.sas`. Notice that is consists of just a `proc format` and a data step. There are no statistical procedures, except a `proc freq` that is commented out. This file will be read by programs that invoke statistical procedures, as you will see.

```
/*******  senicread.sas  Just reads and labels data  ***********/
title 'SENIC data';
options linesize=79;

proc format;  /* value labels used in data step below */
     value yesnofmt 1 = 'Yes'   2 = 'No' ;
     value regfmt  1 =  'Northeast'
                   2 =  'North Central'
                   3 =  'South'
                   4 =  'West' ;
     value acatfmt 1 = '53 & under' 2 = 'Over 53';

data senic;
     infile 'senic.raw' missover ;
            /*  in senic.raw, missing=blank  */
            /*  missover causes all blanks to be missing,
                even at the end of a line. */
     input
         #1  id       1-5
             stay     7-11
             age      13-16
             infrisk  18-20
             culratio 22-25
             xratio   27-31
             nbeds    33-35
             medschl  37
             region   39
             census   41-43
             nurses   45-47
             service  49-52 ;
     label id       = 'Hospital identification number'
           stay     = 'Av length of hospital stay, in days'
           age      = 'Average patient age'
           infrisk  = 'Prob of acquiring infection in hospital'
           culratio = '# cultures  / # no hosp acq infect'
           xratio   = '# x-rays / # no signs of pneumonia'
           nbeds    = 'Average # beds during study period'
           medschl  = 'Medical school affiliation'
           region   = 'Region of country (usa)'
           census   = 'Aver # patients in hospital per day'
           nurses   = 'Aver # nurses during study period'
           service  = '% of 35 potential facil. & services' ;
        /* associating variables with their value labels */
     format medschl yesnofmt.;
     format region  regfmt.;

        /***** recodes, computes & ifs  *****/
```

```
     if 0<age<=53 then agecat=1;
     else if age>53 then agecat=2;
     label   agecat = 'av patient age category';
     format agecat acatfmt.;

  /*  compute ad hoc index of hospital quality  */

     quality=(2*service+nurses+nbeds+10*culratio
                 +10*xratio-2*stay)/medschl;
     if (region eq 3) then quality=quality-100;
     label quality = 'jerry''s bogus hospital quality index';


/* Commented out

proc freq;
      tables _all_;
*/
```

Here are some comments.

- Notice that we are reading the variables from specified columns. This allows data to be packed into adjacent columns (some data files are like this), and also allows missing data to be represented by blanks. But it means that the data must be perfectly aligned into columns. Don't *assume* this is true just because you were told by someone who should know. Check!
- The `missover` option is highly recommended if missing values are represented by blanks.
- `if 0<age<=53` means "if $0 < \text{age} \leq 53$.'
- Age = 0 or negative would result in a missing value for `agecat`.
- a missing value for `xratio` (or any other variable in the formula) would result in a missing value for `quality`.
- The double quotation mark in the middle of the label for `quality` is how you get an apostrophe in a label.
- `tables _all_` in `proc freq`: The reserved name `_all_` means all the variables in the data set.

Here is a program that pulls in `senicread.sas` with a `%include` statement, and then does some statistical tests. Keeping the data definition in a separate file is often a good strategy, because most data analysis projects involve a substantial number of statistical procedures. It is common to have maybe twenty program files that carry out various analyses. You *could* have the data step at the beginning of each program, but what

happens when (inevitably) you want to make a change in the data step and re-run your analyses? You find yourself making the same change in twenty files. Probably you will forget to change some of them, and the result is a big mess. If you keep your data definition in just one place, you only have to edit it once, and a lot of problems are avoided.

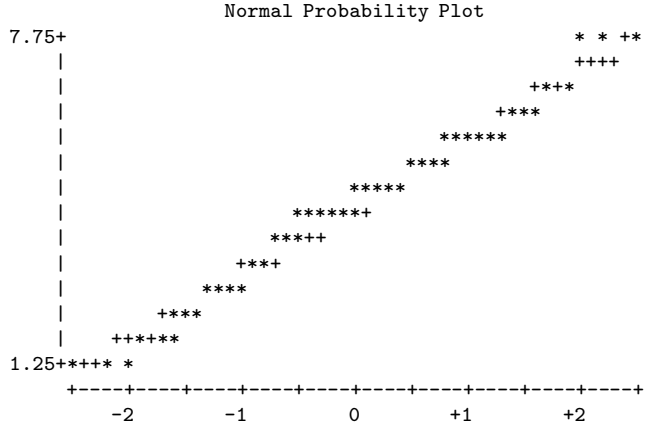
```
/****************** basicsenic.sas ***************/
/*               Basic stats on SENIC Data         */
/*************************************************/

%include 'senicread.sas';  /* senicread.sas reads data, etc.  */

proc univariate  plot normal ; /* Plots and a test for normality */
     title2 'Describe Quantitative Variables';
     var stay -- nbeds census nurses service;
     /* single dash only works with numbered lists, like item1-item50    */
proc freq;
     title2 'Frequency distributions of categorical variables';
     tables medschl region agecat;
proc chart;
     title2 'Vertical bar charts';
     vbar region medschl agecat /discrete ;
proc chart ;
     title2 'Pie chart';
     pie region/type=freq;

proc chart;
     title2 'Pseudo 3-d chart - just playing around';
     block region / sumvar=infrisk type=mean group=medschl discrete;

/* Now elementary tests */

proc freq;  /* use freq to do crosstabs */
     tables region*medschl / nocol nopercent expected chisq;
proc ttest;
     class medschl;
     var infrisk age ;
proc glm;  /* one-way anova */
     class region;
     model infrisk=region;
     means region/ snk scheffe;
proc plot;
     plot infrisk * nurses
          infrisk * nurses = medschl;
proc corr;
     var stay -- nbeds census nurses service;
proc glm;  /*  simple regression with glm*/
```

```
model infrisk=nurses;
```

The list file from this job is long, so we will just look at the `proc univariate` output for the dependent variable.

```
^L                                  SENIC data
6
                        Describe Quantitative Variables
                                                11:47 Friday, January 4, 2002

                             Univariate Procedure

Variable=INFRISK       Prob of acquiring infection in hospital

                                  Moments

                    N              113  Sum Wgts        113
                    Mean       4.354867  Sum           492.1
                    Std Dev    1.340908  Variance   1.798034
                    Skewness   -0.11976  Kurtosis   0.182355
                    USS         2344.41  CSS        201.3798
                    CV         30.79102  Std Mean   0.126142
                    T:Mean=0   34.52353  Pr>|T|       0.0001
                    Num ^= 0        113  Num > 0         113
                    M(Sign)        56.5  Pr>=|M|      0.0001
                    Sgn Rank     3220.5  Pr>=|S|      0.0001
                    W:Normal   0.970897  Pr<W         0.1280

                               Quantiles(Def=5)



                     100% Max       7.8       99%       7.7
                      75% Q3        5.2       95%       6.4
                      50% Med       4.4       90%       5.8
                      25% Q1        3.7       10%       2.6
                       0% Min       1.3        5%       1.8
                                                1%       1.3
                     Range          6.5
                     Q3-Q1          1.5
                     Mode           4.3

                                  Extremes

                       Lowest    Obs     Highest    Obs
                          1.3(      93)      6.5(      47)
                          1.3(      40)      6.6(     104)
                          1.4(     107)      7.6(      53)
                          1.6(       2)      7.7(      13)
                          1.7(      85)      7.8(      54)

^L                                  SENIC data
7
                        Describe Quantitative Variables
                                                11:47 Friday, January 4, 2002

                             Univariate Procedure

Variable=INFRISK       Prob of acquiring infection in hospital

                  Stem Leaf                     #             Boxplot
                     7 678                      3                0
                     7                                           |
                     6 56                       2                |
                     6 12334                    5                |
                     5 5555666777889           13                |
                     5 0000112233344           13             +-----+
                     4 5555555666778888999     19             |     |
                     4 0111122222333333344444  22             *--+--*
```

```
    3 5577778999               10             +-----+
    3 011244                    6                |
    2 5677899999               10                |
    2 0013                      4                |
    1 678                       3                |
    1 334                       3                0
      ----+----+----+----+--
```



2.2.8 SAS Reference Materials

This course is trying to teach you SAS by example, without full explanation, and certainly without discussion of all the options. If you need more detail, there are several approaches you can take. The most obvious is to consult the SAS manuals. The full set of manuals runs to over a dozen volumes, and most of them look like telephone directories. For a beginner, it is hard to know where to start. And even if you know where to look, the SAS manuals can be hard to read, because they assume you already understand the statistical procedures fairly thoroughly, and on a mathematical level. They are really written for professional statisticians. The SAS Institute also publishes a variety of manual-like books that are intended to be more instructional, most of them geared to specific topics (like *The SAS system for multiple regression* and *the SAS system for linear models*). These are a bit more readable, though it helps to have a real textbook on the topic to fill in the gaps.

A better place to start is a wonderful book by Cody and Smith [1] entitled *Applied statistics and the SAS programming language*. They do a really good job of presenting and documenting the language of the data step, and and they also cover a set of statistical procedures ranging from elementary to moderately advanced. If you had to own just one SAS book, this would be it.

If you consult *any* SAS book or manual (Cody and Smith's book included), you'll need to translate and filter out some details. First, you're advised to ignore anything about the SAS Display Manager. In this course, there are raw data file, program files, log files and list files; that's it.

Second, many of the examples you see in Cody and Smith's book and elsewhere will not have separate files for the raw data and the program. They include the raw data in the program file in the data step, after a `datalines` or `cards` statement. Here is an example from page 3 of [1].

```
data test;
```

```
    input subject 1-2 gender $ 4 exam1 6-8 exam2 10-12 hwgrade $ 14;
    datalines;
10 M  80  84 A
 7 M  85  89 A
 4 F  90  86 B
20 M  82  85 B
25 F  94  94 A
14 F  88  84 C
;
proc means data=test;
run;
```

Having the raw data and the SAS code together in one display is so attractive for small datasets that most textbook writers cannot resist it. But think how unpleasant it would be if you had 10,000 lines of data. The way we would do this example is to have the data file (named, say, `example1.dat`) in a separate file. The data file would look like this.

```
10 M  80  84 A
 7 M  85  89 A
 4 F  90  86 B
20 M  82  85 B
25 F  94  94 A
14 F  88  84 C
```

and the program file would look like this.

```
data test;
    infile 'example1.dat'; /* Read data from example1.dat */
    input subject 1-2 gender $ 4 Exam1 6-8 exam2 10-12 hwgrade $ 14;
proc means data=test;
```

Using this as an example, you should be able to translate any textbook example into the program-file data-file format used in this course.

Chapter 3

Multiple Regression: Part One

3.1 Three Meanings of Control

In this class, we will use the word **control** to refer to procedures designed to reduce the influence of extraneous variables on our results. The definition of extraneous is "not properly part of a thing," and we will use it to refer to variables we're not really interested in, and which might get in the way of understanding the relationship between the independent variable and the dependent variable.

There are two ways an extraneous variable might get in the way. First, it could be a confounding variable – related to both the independent variable and the dependent variable, and hence capable of creating masking or even reversing relationships that would otherwise be evident. Second, it could be unrelated to the independent variable and hence not a confounding variable, but it could still have a substantial relationship to the dependent variable. If it is ignored, the variation that it could explain will be part of the "background noise," making it harder to see the relationship between IV and DV, or at least causing it to appear relatively weak, and possibly to be non-significant.

The main way to control potential extraneous variables is by holding them constant. In **experimental control**, extraneous variables are literally held constant by the procedure of data collection or sampling of cases. For example, in a study of problem solving conducted at a high school, background noise might be controlled by doing the experiment at the same time of day for each subject (and not when classes are changing). In learning experiments with rats, males are often employed because their behavior is less variable than that of females.

An alternative to experimental control is **statistical control**, which takes two main forms. One version, **subdivision**, is to subdivide the sample into groups with identical or nearly identical values of the extraneous variable(s), and then to examine the relationship between independent and dependent variable separately in each subgroup – possibly pooling the subgroup analyses in some way. For example, the relationship between education and income might be studied separately for men and women. The drawback of this subdivision approach is that if extraneous variables have many values or combinations of values, you need a very large sample.

The second form of statistical control, **model-based** control, is to exploit details of the statistical model to accomplish the same thing as the subdivision approach, but without needing a huge sample size. The primary example is multiple linear regression, which is

the topic of this chapter.

3.2 Population Parameters

Recall we said two variables are "related" if the distribution of the dependent variable *depends* on the value of the independent variable. Classical regression and analysis of variance are concerned with a particular way in which the independent and dependent variables might be related, one in which the *population mean* of Y depends on the value of X.

Think of a population histogram manufactured out of a thin sheet of metal. The point (along the horizontal axis) where the histogram balances is called the **expected value** or population mean; it is usually denoted by $E[Y]$ or μ (the Greek letter mu). The *conditional* population mean of Y given $X = x$ is just the balance point of the conditional distribution. It will be denoted by $E[Y|X = x]$. The vertical bar — should be read as "given."

Again, for every value of X, there is a separate distribution of Y, and the expected value (population mean) of that distribution depends on the value of X. Furthermore, that dependence takes a very specific and simple form. When there is only one independent variable, the population mean of Y is

$$E[Y|X = x] = \beta_0 + \beta_1 x. \tag{3.1}$$

This is the equation of a straight line. The slope (rise over run) is β_1 and the intercept is β_0. If you want to know the population mean of Y for any given x value, all you need are the two numbers β_0 and β_1.

But in practice, we never know β_0 and β_1. To *estimate* them, we use the slope and intercept of the least-squares line:

$$\widehat{Y} = b_0 + b_1 x. \tag{3.2}$$

If you want to estimate the population mean of Y for any given x value, all you need are the two numbers b_0 and b_1, which are calculated from the sample data.

This has a remarkable implication, one that carries over into multiple regression. Ordinarily, if you want to estimate a population mean, you need a reasonable amount of data. You calculate the sample mean of those data, and that's your estimate of the population mean. If you want to estimate a *conditional* population mean, that is, the population mean of the conditional distribution of Y given a particular $X = x$, you need a healthy amount of data with that value of x. For example, if you want to estimate the average weight of 50 year old women, you need a sample of 50 year old women — unless you are willing to make some assumptions.

What kind of assumptions? Well, the simple structure of (3.1) means that you can use formula (3.2) to estimate the population mean of Y for a given value of $X = x$ *without having any data* at that x value. This is not "cheating," or at any rate, it need not be. If

- the x value in question is comfortably within the range of the data in your sample, and if
- the straight-line model is a reasonable approximation of reality within that range,

then the estimate can be quite good.

The ability to estimate a conditional population mean without a lot of data at any given x value means that we will be able to control for extraneous variables, and remove their influence from a given analysis without having the massive amounts of data required by the subdivision approach to statistical control.

We are getting away with this because we have adopted a *model* for the data that makes reasonably strong assumptions about the way in which the population mean of Y depends on X. If those assumptions are close to the truth, then the conclusions we draw will be reasonable. If the assumptions are badly wrong, we are just playing silly games. There is a general principle here, one that extends far beyond multiple regression.

Data Analysis Hint 3 *There is a direct tradeoff between amount of data and the strength (restrictiveness) of model assumptions. If you have a lot of data, you do not need to assume as much. If you have a small sample size, you will probably have to adopt fairly restrictive assumptions in order to conclude anything from your data.*

Multiple Regression Now consider the more realistic case where there is more than one independent variable. With two independent variables, the model for the population mean of Y is

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

which is the equation of a plane in 3 dimensions (x_1, x_2, y). The general case is

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1},$$

which is the equation of a hyperplane in p dimensions.

Comments

- Since there is more than one independent variable, there is a conditional distribution of Y for every *combination* of independent variable values. Matrix notation (boldface) is being used to denote a collection of independent variables.
- There are $p - 1$ independent variables. This may seem a little strange, but we're doing this to keep the notation consistent with that of standard regression texts such as [3]. If you want to think of an independent variable $X_0 = 1$, then there are p independent variables.
- What is β_0? It's the height of the population hyperplane when all the independent variables are zero, so it's the *intercept*.
- Most regression models have an intercept term, but some do not ($X_0 = 0$); it depends on what you want to accomplish.
- β_0 is the intercept. We will now see that the other β values are slopes.

Consider

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

What is β_3? If you speak calculus, $\frac{\partial}{\partial x_3}E[Y] = \beta_3$, so β_3 is the rate at which the population mean is increasing as a function of x_3, when other independent variables are *held constant* (this is the meaning of a partial derivative).

If you speak high school algebra, β_3 is the change in the population mean of Y when x_3 is increased by one unit and all other independent variables are *held constant*. Look at

$$\begin{array}{rl} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad +\beta_3(x_3+1) \quad +\beta_4 x_4 \\ - & (\beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad +\beta_3 x_3 \qquad\quad +\beta_4 x_4) \\ \\ = & \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad +\beta_3 \quad +\beta_4 x_4 \\ - & \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \qquad\quad -\beta_4 x_4 \\ \\ = & \beta_3 \end{array}$$

The mathematical device of *holding other variables constant* is very important. This is what is meant by statements like "**Controlling for** parents' education, parents' income and number of siblings, quality of day care is still positively related to academic performance in Grade 1." We have just seen the prime example of model-based statistical control — the third type of control in the "Three meanings of control" section that began this chapter.

We will describe the relationship between X_k and Y as **positive** (controlling for the other independent variables) if $\beta_k > 0$ and **negative** if $\beta_k < 0$.

Here is a useful definition. A quantity (say w) is a **linear combination** of quantities z_1, z_2 and z_3 if $w = a_1z_1 + a_2z_2 + a_3z_3$, where a_1, a_2 and a_3 are constants. Common multiple regression is *linear* regression because the population mean of Y is a linear combination of the β values. It does *not* refer to the shape of the curve relating x to $E[Y|X = x]$. For example,

$E[Y\|X = x]$	=	$\beta_0 + \beta_1 x$	Simple linear regression
$E[Y\|X = x]$	=	$\beta_0 + \beta_1 x^2$	Also simple linear regression
$E[Y\|X = x]$	=	$\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$	Polynomial regression – still linear
$E[Y\|X = x]$	=	$\beta_0 + \beta_1 x + \beta_2 \cos(1/x)$	Still linear in the β values
$E[Y\|X = x]$	=	$\beta_0 + \beta_1 \cos(\beta_2 x)$	Truly non-linear

When the relationship between the independent and dependent variables is best represented by a curve, we'll call it **curvilinear**, whether the regression model is linear or not. All the examples just above are curvilinear, except the first one.

Notice that in the polynomial regression example, there is really only one independent variable, x. But in the regression model, x, x^2 and x^3 are considered to be three separate independent variables in a multiple regression. Here, fitting a curve to a cloud of points in two dimensions is accomplished by fitting a hyperplane in four dimensions. The origins of this remarkable trick are lost in the mists of time, but whoever thought of it was having a good day.

3.3 Estimation by least squares

In the last section, the conditional population mean of the dependent variable was modelled as a (population) hyperplane. It is natural to estimate a population hyperplane

with a sample hyperplane. This is easiest to imagine in three dimensions. Think of a three-dimensional scatterplot, in a room. The independent variables are X_1 and X_2. The (x_1, x_2) plane is the floor, and the value of Y is height above the floor. Each subject (case) in the sample contributes three coordinates (x_1, x_2, y), which can be represented by a soap bubble floating in the air.

In simple regression, we have a two-dimensional scatterplot, and we seek the best-fitting straight line. In multiple regression, we have a three (or higher) dimensional scatterplot, and we seek the best fitting plane (or hyperplane). Think of lifting and tilting a piece of plywood until it fits the cloud of bubbles as well as possible.

What is the "best-fitting" plane? We'll use the **least-squares plane**, the one that minimizes the sum of squared vertical distances of the bubbles from the piece of plywood. These vertical distances can be viewed as errors of prediction.

It's hard to visualize in higher dimension, but the algebra is straightforward. Any sample hyperplane may be viewed as an estimate (maybe good, maybe terrible) of the population hyperplane. Following the statistical convention of putting a hat on a population parameter to denote an estimate of it, the equation of a sample hyperplane is

$$\widehat{\beta_0} + \widehat{\beta_1} x_1 + \ldots + \widehat{\beta_{p-1}} x_{p-1},$$

and the error of prediction (vertical distance) is the difference between y and the quantity above. So, the least squares plane must minimize

$$Q = \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \widehat{\beta_1} x_{i,1} - \ldots - \widehat{\beta_{p-1}} x_{i,p-1} \right)^2$$

over all combinations of $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_{p-1}}$.

Provided that no independent variable (including the peculiar $X_0 = 1$) is a perfect linear combination of the others, the $\widehat{\beta}$ quantities that minimize the sum of squares Q exist and are unique. We will denote them by b_0 (the estimate of β_0, b_1 (the estimate of β_1), and so on.

Again, *a population hyperplane is being estimated by a sample hyperplane.*

$$\begin{aligned} E[Y|\boldsymbol{X} = \boldsymbol{x}] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\ \widehat{Y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \end{aligned}$$

- $\widehat{Y}$ means *predicted* Y. It is the height of the best-fitting (least squares) piece of plywood above the floor, at the point represented by the combination of x values. The equation for $\widehat{Y}$ is the equation of the least-squares hyperplane.
- "Fitting the model" means calculating the b values.

3.3.1 Residuals

The **residual**, or error of prediction, is

$$e = Y - \widehat{Y}.$$

The residuals (there are n) represents errors in prediction. A positive residual means over-performance (or under-prediction). A negative residual means under-performance. Examination of residuals can reveal a lot, since we can't look at 12-dimensional scatterplots.

- Single variable plots (histograms, box plots, stem and leaf diagrams etc.) can identify outliers. (Data errors? Source of new ideas? What might a bimodal distribution of residuals indicate?)
- Plot (scatterplot) of residuals versus potential independent variables not in the model might suggest they be included, or not. How would you plot residuals vs a categorical IV?
- Plot of residuals vs. variables that are in the model may reveal
  - Curvilinear trend (may need transformation of x, or polynomial regression, or even real non-linear regression)
  - Non-constant variance over the range of x, so the DV may depend on the IV not just through the mean. May need transformation of Y, or weighted least squares, or a different model.
- Plot of residuals vs. $\widehat{Y}$ may also reveal unequal variance.

3.3.2 Categorical Independent Variables

Independent variables need not be continuous – or even quantitative. For example, suppose subjects in a drug study are randomly assigned to either an active drug or a placebo. Let Y represent response to the drug, and

$$x = \begin{cases} 1 & \text{if the subject received the active drug, or} \\ 0 & \text{if the subject received the placebo.} \end{cases}$$

The model is $E[Y|X = x] = \beta_0 + \beta_1 x$. For subjects who receive the active drug (so $x = 1$), the population mean is

$$\beta_0 + \beta_1 x = \beta_0 + \beta_1$$

For subjects who receive the placebo (so $x = 0$), the population mean is

$$\beta_0 + \beta_1 x = \beta_0.$$

Therefore, β_0 is the population mean response to the placebo, and β_1 is the difference between response to the active drug and response to the placebo. We are very interested in testing whether β_1 is different from zero, and guess what? We get exactly the same t value as from a two-sample t-test, and exactly the same F value as from a one-way ANOVA for two groups.

Exercise Suppose a study has 3 treatment conditions. For example Group 1 gets Drug 1, Group 2 gets Drug 2, and Group 3 gets a placebo, so that the Independent Variable is Group (taking values 1,2,3) and there is some Dependent Variable Y (maybe response to drug again).

Sample Question 3.3.1 *Why is $E[Y|X = x] = \beta_0 + \beta_1 x$ (with x = Group) a silly model?*

Answer to Sample Question 3.3.1 *Designation of the Groups as 1, 2 and 3 is completely arbitrary.*

Sample Question 3.3.2 *Suppose* $x_1 = 1$ *if the subject is in Group 1, and zero otherwise, and* $x_2 = 1$ *if the subject is in Group 2, and zero otherwise, and* $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$*. Fill in the table below.*

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
1			$\mu_1 =$
2			$\mu_2 =$
3			$\mu_3 =$

Answer to Sample Question 3.3.2

Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
1	1	0	$\mu_1 = \beta_0 + \beta_1$
2	0	1	$\mu_2 = \beta_0 + \beta_2$
3	0	0	$\mu_3 = \beta_0$

Sample Question 3.3.3 *What does each* β *value mean?*

Answer to Sample Question 3.3.3 $\beta_0 = \mu_3$*, the population mean response to the placebo.* β_1 *is the difference between mean response to Drug 1 and mean response to the placebo.* β_2 *is the difference between mean response to Drug 21 and mean response to the placebo.*

Sample Question 3.3.4 *Why would it be nice to simultaneously test whether* β_1 *and* β_2 *are different from zero?*

Answer to Sample Question 3.3.4 *This is the same as testing whether all three population means are equal; this is what a one-way ANOVA does. And we get the same F and p values (not really part of the sample answer).*

It is worth noting that all the traditional one-way and higher-way models for analysis of variance and covariance emerge as special cases of multiple regression, with dummy variables representing the categorical independent variables.

More about Dummy Variables The exercise above was based on **indicator dummy variables**, which take a value of 1 for observations where a categorical independent variable takes a particular value, and zero otherwise. Notice that x_1 and x_2 contain the same information as the three-category variable Group. If you know Group, you know x_1 and x_2, and if you know x_1 and x_2, you know Group. In models with an intercept term, a categorical independent variable with k categories is always represented by $k - 1$ dummy variables. If the dummy variables are indicators, the category that does not get an indicator is actually the most important. The intercept is that category's mean, and it is called the **reference category**, because the remaining regression coefficients represent differences between the reference category and the other category. To compare several treatments to a control, make the control group the reference category by *not* giving it an indicator.

Sample Question 3.3.5 *What would happen if you used k indicator dummy variables instead of $k-1$?*

Answer to Sample Question 3.3.5 *The dummy variables would add up to the intercept; the independent variables would be linearly dependent, and the least-squares estimators would not exist.*

Your software might try to save you by throwing one of the dummy variables out, but which one would it discard?

3.3.3 Explained Variation

Before considering any independent variables, there is a certain amount of variation in the dependent variable. The sample mean is the value around which the sum of squared errors of prediction is at a minimum, so it's a least squares estimate of the population mean of Y when there are no independent variables. We will measure the total variation to be explained by the sum of squared deviations around the mean of the dependent variable.

When we do a regression, variation of the data around the least-squares plane represents errors of prediction. It is variation that is *unexplained* by the regression. But it's always less than the variation around the sample mean (Why? Because the least-squares plane could be horizontal). So, the independent variables in the regression have explained *some* of the variation in the dependent variable. Variation in the residuals is variation that is still *unexplained*.

Variation to explain: **Total Sum of Squares**

$$\text{SSTO} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Variation that the regression does not explain: **Error Sum of Squares**

$$\text{SSE} = \sum_{i=1}^{n}(e_i - \overline{e})^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Variation that is explained: **Regression (or Model) Sum of Squares**

$$\text{SSR} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

Regression software (including SAS) displays the sums of squares above in an *analysis of variance summary table*. "Analysis" means to "split up," and that's what we're doing here — splitting up the variation in dependent variable into explained and unexplained parts.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

Variance estimates consist of sums of squares divided by degrees of freedom. "DF" stands for Degrees of Freedom. Sums of squares and degrees of freedom each add up to Total. The F-test is for whether $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ – that is, for whether *any* of the independent variables makes a difference.

The proportion of variation in the dependent variable that is explained by the independent variables (representing *strength of relationship*) is

$$R^2 = \frac{\text{SSR}}{\text{SSTO}}$$

The R^2 from a simple regression is the same as the square of the correlation coefficient: $R^2 = r^2$.

What is a good value of R^2? Well, the weakest relationship I can visually perceive in a scatterplot is around $r = .3$, so I am unimpressed by R^2 values under 0.09. By this criterion, most published results in the social sciences, and many published results in the biological sciences are not strong enough to be scientifically interesting. But this is just my opinion.

3.4 Testing for Statistical Significance in Regression

We are already assuming that there is a separate population defined by each combination of values of the independent variables (the conditional distributions of Y given $\mathbf{X}$), and that the conditional population mean is a linear combination of the β values; the weights of this linear combination are 1 for β_0, and the x values for the other β values. The classical assumptions are that in addition,

- Sample values of Y represent independent observations, conditionally upon the values of the independent variables.
- Each conditional distribution is normal.
- Each conditional distribution has the same population variance.

How important are the assumptions? Well, important for what? The main thing we want to avoid is incorrect p-values, specifically ones that appear smaller than they are – so that we conclude a relationship is present when really we should not. This "Type I error" is very undesirable, because it tends to load the scientific literature with random garbage.

For large samples, the assumption of normality is not important provided no single observation has too much influence. What is meant by a "large" sample? It depends on how severe the violations are. What is "too much" influence? The influence of the most influential observation must tend to zero as the sample size approaches infinity. You're welcome.

The assumption of equal variances can be safely violated provided that the numbers of observations at each combination of IV values are large and close to equal. This is most likely to be the case with designed experiments having categorical independent variables.

The assumption of independent observations is very important, almost always. Examples where this does not hold is if a student takes a test more than once, members of

the same family respond to the same questionnaire about eating habits, litter-mates are used in a study of resistance to cancer in mice, and so on.

When you know in advance which observations form non-independent sets, one option is to average them, and let n be the number of independent sets of observations. There are also ways to incorporate non-independence into the statistical model. We will discuss repeated measures designs, multivariate analysis and other examples later.

3.4.1 The standard F and t-tests

SAS `proc reg` (like other programs) usually starts with an overall F-test, which tests all the independent variables in the equation simultaneously. If this test is significant, we can conclude that one or more of the independent variables is related to the dependent variable.

Again like most programs that do multiple regression, SAS produces t-tests for the individual regression coefficients. If one of these is significant, we can conclude that controlling for all other independent variables in the model, the independent variable in question is related to the dependent variable. That is, each variable is tested controlling for all the others.

It is also possible to test subsets of independent variables, controlling for all the others. For example, in an educational assessment where students use 4 different textbooks, the variable "textbook" would be represented by 3 dummy variables. These variables could be tested simultaneously, controlling for several other variables such as parental education and income, child's past academic performance, experience of teacher, and so on.

In general, to test a subset A of independent variables while controlling for another subset B, fit a model with both sets of variables, and simultaneously test the b coefficients of the variables in subset A; there is an F test for this.

This is 100% equivalent to the following. Fit a model with just the variables in subset B, and calculate R^2. Then fit a second model with the A variables as well as the B variables, and calculate R^2 again. Test whether the increase in R^2 is significant. It's the same F test.

Call the regression model with all the independent variables the **Full Model**, and call the model with fewer independent variables (that is, the model without the variables being tested) the **Reduced Model**. Let SSR_F represent the explained sum of squares from the full model, and SSR_R represent the explained sum of squares from the reduced model.

Sample Question 3.4.1 *Why is $SSR_F \geq SSR_R$?*

Answer to Sample Question 3.4.1 *In the full model, if the best-fitting hyperplane had all the b coefficients corresponding to the extra variables equal to zero, it would fit exactly as well as the hyperplane of the reduced model. It could not do any worse.*

Since $R^2 = \frac{SSR}{SSTO}$, it is clear that $SSR_F \geq SSR_R$ implies that adding independent variables to a regression model can only increase R^2. When these additional independent variables are correlated with independent variables already in the model (as they usually are in an observational study),

- Statistical significance can appear when it was not present originally, because the additional variables reduce error variation, and make estimation and testing more precise.
- Statistical significance that was originally present can disappear, because the new variables explain some of the variation previously attributed to the variables that were significant, so when one controls for the new variables, there is not enough explained variation left to be significant. This is especially true of the t-tests, in which each variable is being controlled for all the others.
- Even the signs of the bs can change, reversing the interpretation of how their variables are related to the dependent variable. This is why it's very important not to leave out important independent variables in an observational study.

The F-test for the full versus reduced model is based on the test statistic

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}, \tag{3.3}$$

where MSE_F is the mean square error for the full model: $MSE_F = \frac{SSE_F}{n-p}$. Equation 3.3 is a very general formula. As we will see, all the standard tests in regression and the usual (fixed effects) Analysis of Variance are special cases of this F-test.

Examples of Full and Reduced Models

At this point, it might help to have some concrete examples. Recall the SENIC data set (catching infection in hospital) that was used to illustrate a collection of elementary significance tests in Section 2.2.7. For reference, here is the `label` statement again.

```
label id       = 'Hospital identification number'
      stay     = 'Av length of hospital stay, in days'
      age      = 'Average patient age'
      infrisk  = 'Prob of acquiring infection in hospital'
      culratio = '# cultures  / # no hosp acq infect'
      xratio   = '# x-rays / # no signs of pneumonia'
      nbeds    = 'Average # beds during study period'
      medschl  = 'Medical school affiliation'
      region   = 'Region of country (usa)'
      census   = 'Aver # patients in hospital per day'
      nurses   = 'Aver # nurses during study period'
      service  = '% of 35 potential facil. & services' ;
```

The SAS program `senicread.sas` could have defined dummy variables for `region` and `medschl` in the data step as follows:

```
if region = 1 then r1=1; else r1=0;
if region = 2 then r2=1; else r2=0;
if region = 3 then r3=1; else r3=0;
if medschl = 2 then mschool = 0; else mschool = medschl;
/* mschool is an indicator for medical school = yes */
```

The definition of `r1, r2` and `r2` above is correct, but it is risky. It works only because the data file happens to have no missing values for region. If there were missing values for region, the `else` statements would assign them to zero for `r1, r2` and `r3`, because `else` means *anything* else. The definition of `mschool` is a bit more sophisticated; missing values for `medschl` will also be missing for `mschool`.

Here is what I'd recommend for `region`. It's more trouble, but it's worth it.

```
/* Indicator dummy variables for region */
if region = . then r1=.;
   else if region = 1 then r1 = 1;
   else r1 = 0;
if region = . then r2=.;
   else if region = 2 then r2 = 1;
   else r2 = 0;
if region = . then r3=.;
   else if region = 3 then r3 = 1;
   else r3 = 0;
```

When you create dummy variables with `if` statements, always do crosstabulations of the new dummy variables by the categorical variable they represent, to make sure you did it right. Use the option of `proc freq` to see what happened to the missing values (`missprint` makes "missing" a value of the variables).

```
proc freq;
     tables region * (r1-r3) / missprint nocol norow nopercent ;
```

Sample Question 3.4.2 *Controlling for hospital size as represented by number of beds and number of patients, is average patient age related to infection risk?*

1. *What are the variables in the full model?*
2. *What are the variables in the reduced model?*

Answer to Sample Question 3.4.2

1. `nbeds, census, age`
2. `nbeds, census`

I would never ask for SAS syntax on a test, but for completeness,

```
proc reg;
    model infrisk = nbeds, census, age;
    size: test age=0;
```

Sample Question 3.4.3 *Controlling for average patient age and hospital size as represented by number of beds and number of patients, does infection risk differ by region of the country?*

1. *What are the variables in the full model?*

2. *What are the variables in the reduced model?*

Answer to Sample Question 3.4.3

1. `age, nbeds, census, r1, r2, r3`
2. `age, nbeds, census`

To test the full model versus the reduced model,

```
proc reg;
    model infrisk = age nbeds census r1 r2 r3;
    regn: test r1=r2=r3=0;
```

Sample Question 3.4.4 *Controlling for number of beds, number of patients, average length of stay and region of the country, are number of nurses and medical school affiliation (considered simultaneously) significant predictors of infection risk?*

1. *What are the variables in the full model?*
2. *What are the variables in the reduced model?*

Answer to Sample Question 3.4.4

1. `nbeds, census, stay, r1, r2, r3, nurses, mschool`
2. `nbeds, census, stay, r1, r2, r3`

To test the full model versus the reduced model,

```
proc reg;
    model infrisk = nbeds census stay r1 r2 r3 nurses mschool;
    nursmeds: test nurses=mschool=0;
```

Sample Question 3.4.5 *Controlling for average age of patient, average length of stay and region of the country, is hospital size (as represented by number of beds and number of patients) related to infection risk?*

1. *What are the variables in the full model?*
2. *What are the variables in the reduced model?*

Answer to Sample Question 3.4.5

1. `age, stay, r1, r2, r3, nbeds, census`
2. `age, stay, r1, r2, r3`

To test the full model versus the reduced model,

```
proc reg;
    model infrisk = nbeds census stay r1 r2 r3 nurses mschool;
    size2: test nurses=mschool=0;
```

Sample Question 3.4.6 *Controlling for region of the country and medical school affiliation, are average length of stay and average patient age (considered simultaneously) related to infection risk?*

1. *What are the variables in the full model?*
2. *What are the variables in the reduced model?*

Answer to Sample Question 3.4.6

1. `r1, r2, r3, mschool, stay age`
2. `r1, r2, r3, mschool`

To test the full model versus the reduced model,

```
proc reg;
    model infrisk = nbeds census stay r1 r2 r3 nurses mschool;
    agestay: test age=stay=0;
```

The pattern should be clear. You are "controlling for" the variables in the reduced model. You are testing for the additional variables that appear in the full model but not the reduced model.

Looking at the Formula for F

Formula 3.3 reveals some important properties of the F-test. Bear in mind that the p-value is the area under the F-distribution curve *above* the value of the F statistic. Therefore, anything that makes the F statistic bigger will make the p-value smaller, and if it is small enough, the results will be significant. And significant results are what we want, if in fact the full model is closer to the truth than the reduced model.

- Since there are s more variables in the full model than in the reduced model, the numerator of (3.3) is the *average* improvement in explained sum of squares when we compare the full model to the reduced model. Thus, some of the extra variables might be useless for prediction, but the test could still be significant at least one of them contributes a lot to the explained sum of squares, so that the *average* increase is substantially more than one would expect by chance.
- On the other hand, useless extra independent variables can dilute the contribution of extra independent variables with modest but real explanatory power.
- The denominator is a variance estimate based on how spread out the residuals are. The smaller this denominator is, the larger the F statistic is, and the more likely it is to be significant. Therefore, *control* extraneous sources of variation.
  - If possible, always collect data on any potential independent variable that is known to have a strong relationship to the dependent variable, and include it in both the full model and the reduced model. This will make the analysis more sensitive, because increasing the explained sum of squares will reduce the unexplained sum of squares. You will be more likely to detect a real result as significant, because it will be more likely to show up against the reduced background noise.

- On the other hand, the denominator of formula (3.3) for F is $MSE_F = \frac{SSE_F}{n-p}$, where the number of independent variables is $p-1$. Adding useless independent variables to the model will increase the explained sum of squares by at least a little, but the denominator of MSE_F will go down by one, making MSE_F bigger, and F smaller. The smaller the sample size n, the worse the effect of useless independent variables. You have to be selective.
- The (internal) validity of most experimental research depends on experimental designs and procedures that balance sources of extraneous variation evenly across treatments. But even better are careful experimental procedures that eliminate random noise altogether, or at least hold it to very low levels. Reduce sources of random variation, and the residuals will be smaller. The MSE_F will be smaller, and F will be bigger if something is really going on.
- Most dependent variables are just indirect reflections of what the investigator would really like to study, and in designing their studies, scientists routinely make decisions that are tradeoffs between expense (or convenience) and data quality. When dependent variables represent low-quality measurement, they essentially contain random variation that cannot be explained. This variation will show up in the denominator of (3.3), reducing the chance of detecting real results against the background noise. An example of a dependent variable that might have too much noise would be a questionnaire or subscale of a questionnaire with just a few items.

The comments above sneaked in the topic of **statistical power** by discussing the formula for the F-test. Statistical power is *the probability of getting significant results when something is really going on in the population.* It should be clear that high power is good. We have just seen that statistical power can be increased by including important explanatory variables in the study, by carefully controlled experimental conditions, and by quality measurement. Power can also be increased by increasing the sample size. All this is true in general, and does not depend on the use of the traditional F test.

3.4.2 Connections between Explained Variation and Significance Testing

If you divide numerator and denominator of Equation (3.3) by $SSTO$, the numerator becomes $(R^2_F - R^2_R)/s$, so we see that the F test is based on change in R^2 when one moves from the reduced model to the full model. But the F test for the extra variables (controlling for the ones in the reduced model) is based not just on $R^2_F - R^2_R$, but on a quantity I'll denote by $a = \frac{R^2_F - R^2_R}{1-R^2_R}$. This expresses change in R^2 as a *proportion* of the variation left unexplained by the reduced model. That is, it's the *proportion of remaining variation* that the additional variables explain.

This is actually a more informative quantity than simple change in R^2. For example, suppose you're controlling for a set of variables that explain 80% of the variation in the dependent variable, and you test a variable that accounts for an additional 5%. You have explained 25% of the remaining variation – much more impressive than 5%.

The a notation is non-standard. It's sometimes called a squared multiple partial correlation, but the usual notation for partial correlations is intricate and hard to look

at, so we'll just use a.

You may recall that an F test has two degree of freedom values, a numerator degrees of freedom and a denominator degrees of freedom. In the F test for a full versus reduced model, the numerator degrees of freedom is s, the number of extra variables. The denominator degrees of freedom is $n - p$. Recall that the sample size is n, and if the regression model has an intercept, there are $p - 1$ independent variables. Applying a bit of high school algebra to Equation (3.3), we see that the relationship between F and a is

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right). \tag{3.4}$$

so that for any given sample size, the bigger a becomes, the bigger F is. Also, for a given value of $a \neq 0$, F increases as a function of n. This means you can get a large F (and if it's large enough it will be significant) from strong results and a small sample, *or* from weak results and a large sample. Again, examining the formula for the F statistic yields a valuable insight.

Expression (3.4) for F can be turned around to express a in terms of F, as follows:

$$a = \frac{sF}{n - p + sF} \tag{3.5}$$

This is a useful formula, because scientific journals often report just F values, degrees of freedom and p-values. It's easy to tell whether the results are significant, but not whether the results are strong in the sense of explained variation. But the equality (3.5) above lets you recover information about strength of relationship from the F statistic and its degrees of freedom. For example, based on a three-way ANOVA where the dependent variable is rot in potatoes, suppose the authors write "The interaction of bacteria by temperature was just barely significant (F=3.26, df=2,36, p=0.05)." What we want to know is, once one controls for other effects in the model, what proportion of the remaining variation is explained by the temperature-by-bacteria interaction?

We have s=2, $n - p = 36$, and $a = \frac{2\times 3.26}{36+(2\times 3.26)} = 0.153$. So this effect is explaining a respectable 15% of the variation that remains after controlling for all the other main effects and interactions in the model.

3.5 Multiple Regression with SAS

It is always good to start with a textbook example, so that interested students can locate a more technical discussion of what is going on. The following example is based on the "Dwaine Studios" Example from Chapter 6 of [3]. The observations correspond to photographic portrait studios in 21 towns. In addition to sales (the dependent variable), the data file contains number of children 16 and younger in the community (in thousands of persons), and per capita disposable income in thousands of dollars. Here is the SAS program.

```
/* appdwaine1.sas */
options linesize=79;
title 'Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al';
title2 'Just the defaults';
```

```
data portrait;
     infile 'dwaine.dat';
     input kids income sales;
proc reg;
     model sales = kids income;
/*   model DV(s) = IV(s);          */
```

Here is the list file `appdwaine1.lst`.

```
        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      1
                              Just the defaults 15:52 Sunday, January 13, 2002

Model: MODEL1
Dependent Variable: SALES

                              Analysis of Variance

                                 Sum of          Mean
      Source            DF      Squares        Square      F Value       Prob>F

      Model              2  24015.28211  12007.64106       99.103       0.0001
      Error             18   2180.92741    121.16263
      C Total           20  26196.20952

          Root MSE      11.00739     R-square       0.9167
          Dep Mean     181.90476     Adj R-sq       0.9075
          C.V.           6.05118

                              Parameter Estimates

                    Parameter      Standard    T for H0:
     Variable  DF      Estimate         Error   Parameter=0    Prob > |T|

     INTERCEP   1    -68.857073   60.01695322        -1.147        0.2663
     KIDS       1      1.454560    0.21178175         6.868        0.0001
     INCOME     1      9.365500    4.06395814         2.305        0.0333
```

Here are some comments on the list file.

- First the ANOVA summary table for the overall F-test, testing all the independent variables simultaneously. In `C Total`, `C` means corrected for the sample mean. The p-value of 0.0001 actually means $p < 0.0001$, in this version of SAS. It's better in later versions.
- `Root MSE` is the square root of MSE.
- `Dep Mean` is the mean of the dependent variable.
- `C.V.` is the coefficient of variation – the standard deviation divided by the mean. Who cares?
- `R-square` is R^2
- `Adj R-sq`: Since R^2 never goes down when you add independent variables, models with more variables always look as if they are doing better. Adjusted R^2 is an attempt to penalize the usual R^2 for the number of independent variables in the model. It can be useful if you are trying to compare the predictive usefulness of models with different numbers of variables.
- `Parameter Estimates` are the b values. `Standard Error` is the (estimated) standard deviation of the sampling distribution of b. It's the denominator of the t test in the next column.
- The last column is a two-tailed p-value for the t-test.

Here are some sample questions based on the list file.

Sample Question 3.5.1 *Suppose we wish to test simultaneously whether number of kids 16 and under and average family income have any relationship to sales. Give the value of the test statistic, and the associated p-value.*

Answer to Sample Question 3.5.1 $F = 99.103$, $p < 0.0001$

Sample Question 3.5.2 *What can you conclude from just this one test?*

Answer to Sample Question 3.5.2 *Sales is related to either number of kids 16 and under, or average family income, or both. But you'd never do this. You have to look at the rest of the printout to tell what's happening.*

Sample Question 3.5.3 *What percent of the variation in sales is explained by number of kids 16 and under and average family income?*

Answer to Sample Question 3.5.3 *91.67%*

Sample Question 3.5.4 *Controlling for average family income, is number of kids 16 and under related to sales?*

1. *What is the value of the test statistic?*
2. *What is the p-value?*
3. *Are the results significant? Answer Yes or No.*
4. *Is the relationship positive, or negative?*

Answer to Sample Question 3.5.4

1. $t = 6.868$
2. $p < 0.0001$
3. Yes.
4. Positive.

Sample Question 3.5.5 *Controlling for number of kids 16 and under is average family income related to sales?*

1. *What is the value of the test statistic?*
2. *What is the p-value?*
3. *Are the results significant? Answer Yes or No.*
4. *Is the relationship positive, or negative?*

Answer to Sample Question 3.5.5

1. $t = 2.305$

2. $p = 0.0333$

3. Yes.

4. Positive.

Sample Question 3.5.6 *What do you conclude from this entire analysis? Direct your answer to a statistician or researcher.*

Answer to Sample Question 3.5.6 *Number of kids 16 and under and average family income are both related to sales, even when each variable is controlled for the other.*

Sample Question 3.5.7 *What do you conclude from this entire analysis? Direct your answer to a person without statistical training.*

Answer to Sample Question 3.5.7 *Even when you allow for the number of kids 16 and under in a town, the higher the average family income in the town, the higher the average sales. When you allow for the average family income in a town, the higher the number of children under 16, the higher the average sales.*

Sample Question 3.5.8 *A new studio is to be opened in a town with 65,400 children 16 and under, and an average household income of $17,600. What annual sales do you predict?*

Answer to Sample Question 3.5.8 $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2$ *= -68.857073 + 1.454560*65.4 + 9.365500*17.6 = 191.104, so predicted annual sales = $191,104.*

Sample Question 3.5.9 *For any fixed value of average income, what happens to predicted annual sales when the number of children under 16 increases by one thousand?*

Answer to Sample Question 3.5.9 *Predicted annual sales goes up by $1,454.*

Sample Question 3.5.10 *What do you conclude from the t-test for the intercept?*

Answer to Sample Question 3.5.10 *Nothing. Who cares if annual sales equals zero for towns with no children under 16 and an average household income of zero?*

The final two questions ask for a proportion of remaining variation, the quantity we are denoting by a. If you were doing an analysis yourself and wanted this statistic, you'd likely fit a full and a reduced model (or obtain sequential sums of squares; we'll see how to do this in the next example), and calculate the answer directly. But in the published literature, sometimes all you have are reports of t-tests for regression coefficients.

Sample Question 3.5.11 *Controlling for average household income, what proportion of the remaining variation is explained by number of children under 16?*

Answer to Sample Question 3.5.11 *Using* $F = t^2$ *and plugging into (3.5), we have* $a = \frac{1 \times 6.868^2}{21-3+1 \times 6.868^2} = 0.691944$*, or around 70% of the remaining variation.*

Sample Question 3.5.12 *Controlling for number of children under 16, what proportion of the remaining variation is explained by average household income?*

Answer to Sample Question 3.5.12 $a = \frac{2.305^2}{18+2.305^2} = 0.2278994$*, or about 23%.*

These a values are large, but the sample size is small; after all, it's a textbook example, not real data. Now here is a program file that illustrates some options, and gives you a hint of what a powerful tool SAS can be.

```
/* appdwaine2.sas */
options linesize=79 pagesize=35;
title 'Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al';
title2 'With bells and whistles';

data portrait;
     infile 'dwaine.dat';
     input kids income sales;

proc reg simple corr;     /* "simple" prints simple descriptive statistics */
     model sales = kids income / ss1;     /* "ss1" prints Sequential SS */
     output out=resdata predicted=presale  residual=resale;
     /* Creates new SAS data set with Y-hat and e as additional variables*/
     /* Now all the default F-test, in order */
           allivs: test kids = 0, income = 0;
           inter:  test intercept=0;
           child:  test kids=0;
           money:  test income=0;

proc iml; /* Income controlling for kids: Full vs reduced by "hand" */
     fcrit = finv(.95,1,18); print fcrit;
     /* Had to look at printout from an earlier run to get these numbers*/
     f = 643.475809 / 121.16263; /* Using the first F formula */
     pval = 1-probf(f,1,18);
     tsq = 2.305**2; /* t-squared should equal F*/
     a = 643.475809/(26196.20952 - 23372);
     print f tsq pval;
     print "Proportion of remaining variation is " a;

proc glm; /* Use proc glm to get a y-hat more easily */
     model sales=kids income;
     estimate 'Xh p249'  intercept 1   kids 65.4   income 17.6;

proc print; /* To see the new data set with residuals*/
proc univariate normal plot;
     var resale;
proc plot;
     plot resale * (kids income sales);
```

Here are some comments on `appdwaine2.sas`.

- `simple corr` You could get means and standard deviations from `proc means` and correlations from `proc corr`, but this is convenient.
- `ss1` These are Type I Sums of Squares, produced by default in `proc glm`. In `proc reg`, you must request them is you want to see them. The independent variables in the `model` statement are added to the model in order, so that for each variable, the reduced model has all the variables that come before it, and the full model has all

those variables *plus* the current one. The `ss1` option shows the *increase* in explained sum of squares that comes from adding each variable to the model, in the order they appear in the `model` statement.

- `output` creates a new sas data set called `resdata`. It has all the variables in the data set `portrait`, and in addition it has $\widehat{Y}$ (named `presale` for predicted sales) and e (named `resale` for residual of sales).

- Then we have some custom tests, all of them equivalent to what we would get by testing a full versus reduced model. SAS takes the approach of testing whether s linear combinations of β values equal s specified constants (usually zero). Again, this is the same thing as testing a full versus a reduced model. The form of a custom test in `proc reg` is

  1. A name for the test, 8 characters or less, followed by a colon; this name will be used to label the output.
  2. the word `test`.
  3. s linear combinations of independent variable names, each set equal to some constant, separated by commas.
  4. A semi-colon to end, as usual.

  If you want to think of the significance test in terms of a collection of linear combinations that specify constraints on the β values (this is what a statistician would appreciate), then we would say that the names of the independent variables (including the weird variable "intercept") are being used to refer to the corresponding βs. But usually, you are testing a subset of independent variables controlling for some other subset. In this case, include all the variables in the `model` statement, and set the variables you are testing equal to zero in the `test` statement. Commas are optional. As an example, for the test `allivs` (all independent variables) we could have written `allivs: test kids = income = 0;`.

- Now suppose you wanted to use the Sequential Sums of Squares to test `income` controlling for `kids`. You could use a calculator and a table of the F distribution from a textbook, but for larger sample sizes the exact denominator degrees of freedom you need are seldom in the table, and you have to interpolate in the table. With `proc iml` (Interactive Matrix Language), which is actually a nice programming environment, you can use SAS as your calculator. Among other things, you can get exact critical values and p-values quite easily. Statistical tables are obsolete.

  In this example, we first get the **critical value** for F; *if the test statistic is bigger than the critical value, the result is significant.* Then we calculate F using formula 3.3 and its p-value. This F should be equal to the square of the t statistic from the printout, so we check. Then we use (3.5) to calculate a, and print the results.

- `proc glm` The `glm` procedure is very useful when you have categorical independent variables, because it makes your dummy variables for you. But it also can do multiple regression. This example calls attention to the `estimate` command, which lets you calculate $\widehat{Y}$ values more easily and with less chance of error than with a calculator or `proc iml`.

- `proc print` prints all the data values, for all the variables. This is a small data set, so it's not producing a telephone book here. You can limit the variables and the number of cases it prints; see the manual or *Applied statistics and the SAS programming language* [1]. By default, all SAS procedures use the most recently created SAS data set; this is `resdata`, which was created by `proc reg` – so the predicted values and residuals will be printed by `proc print`.

- You didn't notice, but `proc glm` also used `resdata` rather than `portrait`. But it was okay, because `resdata` has all the variables in `portrait`, and *also* the predicted Y and the residuals.

- `proc univariate` produces a lot of useful descriptive statistics, along with a fair amount of junk. The `normal` option gives some tests for normality, and textttplot generates some line-printer plots like boxplots and stem-and-leaf displays. These are sometimes informative. It's a good idea to run the residuals (from the full model) through `proc univariate` if you're starting to take an analysis seriously.

- `proc plot` This is how you would plot residuals against variables in the model. It the data file had additional variables you were *thinking* of including in the analysis, you could plot them against the residuals too, and look for a correlation. My personal preference is to start plotting residuals fairly late in the exploratory game, once I am starting to get attached to a regression model.

Here is the list file `appdwaine2.lst`.

```
 Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      1
                        With bells and whistles
                                        10:58 Saturday, January 19, 2002

                        Descriptive Statistics

 Variables                 Sum               Mean      Uncorrected SS

 INTERCEP                   21                  1                  21
 KIDS                     1302.4        62.019047619            87707.94
 INCOME                     360        17.142857143             6190.26
 SALES                     3820         181.9047619            721072.4


          Variables            Variance       Std Deviation

          INTERCEP                    0                  0
          KIDS             346.71661905        18.620328113
          INCOME           0.9415714286        0.9703460355
          SALES            1309.8104762        36.191303875


                              Correlation

        CORR                KIDS            INCOME             SALES

        KIDS              1.0000            0.7813            0.9446
        INCOME            0.7813            1.0000            0.8358
        SALES             0.9446            0.8358            1.0000
^L      Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al
2
                        With bells and whistles
                                        10:58 Saturday, January 19, 2002

Model: MODEL1
```

Dependent Variable: SALES

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	24015.28211	12007.64106	99.103	0.0001
Error	18	2180.92741	121.16263		
C Total	20	26196.20952			

Root MSE	11.00739	R-square	0.9167
Dep Mean	181.90476	Adj R-sq	0.9075
C.V.	6.05118		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-68.857073	60.01695322	-1.147	0.2663
KIDS	1	1.454560	0.21178175	6.868	0.0001
INCOME	1	9.365500	4.06395814	2.305	0.0333

Variable	DF	Type I SS
INTERCEP	1	694876
KIDS	1	23372
INCOME	1	643.475809

Dependent Variable: SALES
Test: ALLIVS Numerator: 12007.6411 DF: 2 F value: 99.1035
Denominator: 121.1626 DF: 18 Prob>F: 0.0001

Dependent Variable: SALES
Test: INTER Numerator: 159.4843 DF: 1 F value: 1.3163
Denominator: 121.1626 DF: 18 Prob>F: 0.2663

Dependent Variable: SALES
Test: CHILD Numerator: 5715.5058 DF: 1 F value: 47.1722
Denominator: 121.1626 DF: 18 Prob>F: 0.0001

Dependent Variable: SALES
Test: MONEY Numerator: 643.4758 DF: 1 F value: 5.3108
Denominator: 121.1626 DF: 18 Prob>F: 0.0333

FCRIT
4.4138734

F TSQ PVAL
5.3108439 5.313025 0.0333214

A
Proportion of remaining variation is 0.2278428

General Linear Models Procedure

Number of observations in data set = 21

General Linear Models Procedure

Dependent Variable: SALES

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24015.282112	12007.641056	99.10	0.0001
Error	18	2180.927411	121.162634		
Corrected Total	20	26196.209524			

R-Square	C.V.	Root MSE	SALES Mean
0.916746	6.051183	11.007390	181.90476

Source	DF	Type I SS	Mean Square	F Value	Pr > F
KIDS	1	23371.806303	23371.806303	192.90	0.0001
INCOME	1	643.475809	643.475809	5.31	0.0333

Source	DF	Type III SS	Mean Square	F Value	Pr > F
KIDS	1	5715.5058347	5715.5058347	47.17	0.0001
INCOME	1	643.4758090	643.4758090	5.31	0.0333

General Linear Models Procedure

Dependent Variable: SALES

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
Xh p249	191.103930	69.07	0.0001	2.76679783

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	-68.85707315	-1.15	0.2663	60.01695322
KIDS	1.45455958	6.87	0.0001	0.21178175
INCOME	9.36550038	2.30	0.0333	4.06395814

OBS	KIDS	INCOME	SALES	PRESALE	RESALE
1	68.5	16.7	174.4	187.184	-12.7841
2	45.2	16.8	164.4	154.229	10.1706
3	91.3	18.2	244.2	234.396	9.8037

```
                4    47.8     16.3     154.6    153.329      1.2715
                5    46.9     17.3     181.6    161.385     20.2151
                6    66.1     18.2     207.5    197.741      9.7586
                7    49.5     15.9     152.8    152.055      0.7449
                8    52.0     17.2     163.2    167.867     -4.6666
                9    48.9     16.6     145.4    157.738    -12.3382
               10    38.4     16.0     137.2    136.846      0.3540
               11    87.9     18.3     241.9    230.387     11.5126
               12    72.8     17.1     191.1    197.185     -6.0849
               13    88.4     17.4     232.0    222.686      9.3143
               14    42.9     15.8     145.3    141.518      3.7816
               15    52.5     17.8     161.1    174.213    -13.1132
               16    85.7     18.4     209.7    228.124    -18.4239
               17    41.3     16.5     146.4    145.747      0.6530
               18    51.7     16.3     144.0    159.001    -15.0013
               19    89.6     18.1     232.6    230.987      1.6130
               20    82.7     19.1     224.1    230.316     -6.2161
               21    52.3     16.0     166.5    157.064      9.4356
       Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      9
                           With bells and whistles
9
                           With bells and whistles
                                        11:41 Saturday, January 19, 2002

                            Univariate Procedure

Variable=RESALE        Residual

                                  Moments

                    N               21  Sum Wgts         21
                    Mean             0  Sum               0
                    Std Dev   10.44253  Variance    109.0464
                    Skewness  -0.09705  Kurtosis    -0.79427
                    USS       2180.927  CSS         2180.927
                    CV               .  Std Mean    2.278746
                    T:Mean=0         0  Pr>|T|       1.0000
                    Num ^= 0        21  Num > 0          13
                    M(Sign)        2.5  Pr>=|M|      0.3833
                    Sgn Rank       1.5  Pr>=|S|      0.9599
                    W:Normal  0.955277  Pr<W         0.4190

                               Quantiles(Def=5)

                     100% Max  20.21507       99%  20.21507
                      75% Q3   9.435601       95%  11.51263
                      50% Med  0.744918       90%  10.17057
                      25% Q1   -6.21606       10%  -13.1132
                       0% Min  -18.4239        5%  -15.0013
                                               1%  -18.4239
                     Range     38.63896
                     Q3-Q1     15.65166
                     Mode      -18.4239

^L      Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al     1
0
                           With bells and whistles
                                        11:41 Saturday, January 19, 2002

                            Univariate Procedure

Variable=RESALE        Residual

                                  Extremes

                       Lowest    Obs     Highest    Obs
                     -18.4239(     16) 9.758578(      6)
                     -15.0013(     18) 9.803676(      3)
                     -13.1132(     15) 10.17057(      2)
                     -12.7841(      1) 11.51263(     11)
                     -12.3382(      9) 20.21507(      5)
```

```
Stem Leaf                     #             Boxplot
   2 0                        1                |
   1                                           |
   1 0002                     4                |
   0 99                       2             +-----+
   0 011124                   6             *--+--*
  -0                                         |     |
  -0 665                      3             +-----+
  -1 332                      3                |
  -1 85                       2                |
     ----+----+----+----+
  Multiply Stem.Leaf by 10**+1
```


1

With bells and whistles
11:41 Saturday, January 19, 2002

Univariate Procedure

Variable=RESALE Residual

Normal Probability Plot

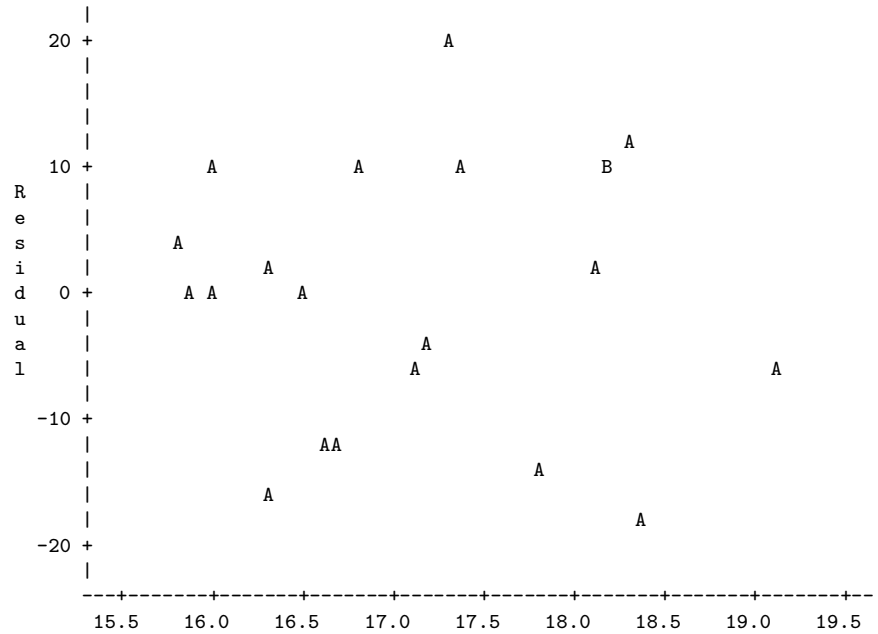
```
            22.5+                                          *****
                |                                      +++++
                |                                  +++++
                |                              ******
             2.5+                       *******
                |                   *++++
                |              +++++
                |          +++++ *
            -17.5+      ******
                  +----+----+----+----+----+----+----+----+----+----+
                      -2        -1         0        +1        +2
```


With bells and whistles
11:32 Tuesday, January 15, 2002

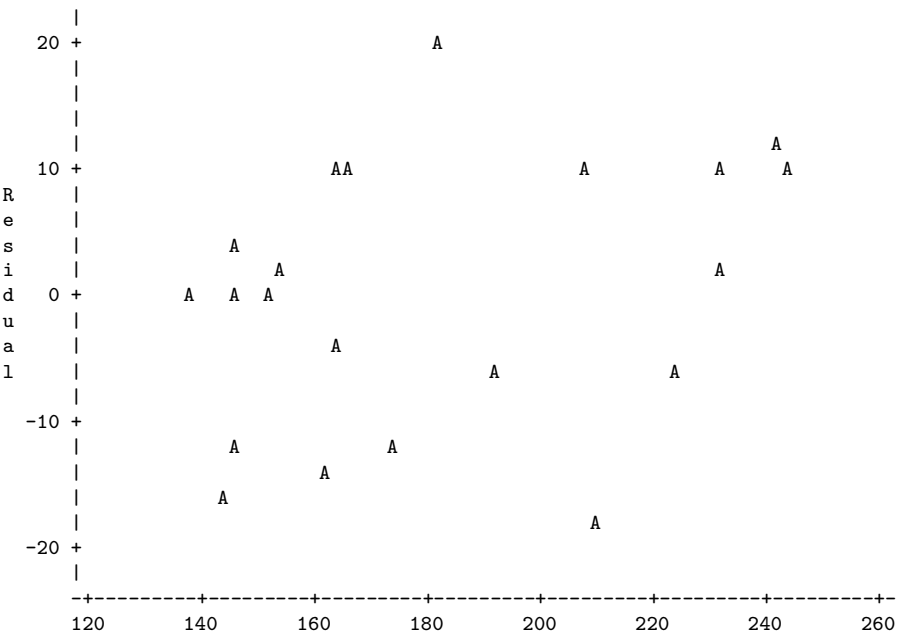
Plot of RESALE*KIDS. Legend: A = 1 obs, B = 2 obs, etc.

```
     |
   20 +                A
     |
     |
     |
     |                                                    A
   10 +              A      A             A                A  A
R    |
e    |
s    |            A
i    |                A                                     A
d  0 +        A  A        A
u    |
a    |                     A
l    |                                    A         A
     |
  -10 +
     |                  A                 A
     |                      A
     |                     A
     |                                                 A
  -20 +
     |
     -+---------+---------+---------+---------+---------+---------+---------+-
     30        40        50        60        70        80        90        100

                                    KIDS
```

Plot of RESALE*INCOME. Legend: A = 1 obs, B = 2 obs, etc.



Plot of RESALE*SALES. Legend: A = 1 obs, B = 2 obs, etc.



Here are some comments.

- `proc reg`
  - In the descriptive statistics produced by the `simple` option, one of the "variables" is `INTERCEP`; it's our friend $X_0 = 1$. The SAS programmers (or the statisticians directing them) are really thinking of this as an independent variable.
  - The Type I (sequential) sum of squares starts with `INTERCEP`, and a really big number for the explained sum of squares. Well, think of a reduced model that does not even have an intercept — that is, one in which there are not only no independent variables, but the population mean is zero. Then add an intercept, so the full model is $E[Y] = \beta_0$. The least squares estimate of β_0 is $\overline{Y}$, so the improvement in explained sum of squares is $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = SSTO$. That's the first line. It makes sense, in a twisted way.
  - Then we have the custom tests, which reproduce the default tests, in order. See how useful the *names* of the custom tests can be?
- `proc iml`: Everything works as advertised. $F = t^2$ except for rounding error, and a is exactly what we got as the answer to Sample Question 3.5.12.
- `proc glm`
  - After an overall test, we get tests labelled `Type I SS` and `Type III SS`. As mentioned earlier, Type One sums of squares are sequential. Each variable is added in turn to the model, in the order specified by the model statement. Each one is tested controlling for the ones that precede it.
  - When independent variables are correlated with each other and with the dependent variable, some of the variation in the dependent variable is being explained by the variation *shared* by the correlated independent variables. Which one should get credit? If you use sequential sums of squares, the variable named first *by you* gets all the credit. And your conclusions can change radically as a result of the order in which you name the independent variables. This may be okay, if you have strong reasons for testing A controlling for B and not the other way around.

    In Type Three sums of squares, each variable is controlled for *all* the others. This way, nobody gets credit for the overlap. It's conservative, and valuable. Naturally, the last lines of Type I and Type III summary tables are identical, because in both cases, the last variable named is being controlled for all the others.
  - I can never remember what Type II and Type IV sums of squares are.
  - The `estimate` statement yielded an `Estimate`, that is, a $|widehatY$ value, of 191.103930, which is what we got with a calculator as the answer to Sample Question 3.5.8. We also get a t-test for whether this particular linear combination differs significantly from zero — insane in this particular case, but useful at other times. The standard error would be very useful if we were constructing confidence intervals or prediction intervals around the estimate, but we are not.

  - Then we get a display of the b values and associated t-tests, as in `proc reg`. I believe these are produced by `proc glm` only when none of the independent variables is declared categorical with the `class` statement.

- `proc print` output is self-explanatory. If you are using `proc print` to print a large number of cases, consider specifying a large page size in the `options` statement. Then, the *logical* page length will be very long, as if you were printing on a long roll of paper, and SAS will not print a new page header with the date and title and so on every 24 line or 35 lines or whatever.

- `proc univariate`: There is so much output to explain, I almost can't stand it. I'll do most of it in class, and just hit a few high points here.

  - `T:Mean=0` A t-test for whether the mean is zero. If the variable consisted of difference scores, this would be a matched t-test. Here, because the mean of residuals from a multiple regression is *always* zero as a by-product of least-squares, t is exactly zero and the p-value is exactly one.
  - `M(Sign)` Sign test, a non-parametric equivalent to the matched t.
  - `Sgn Rank` Wilcoxon's signed rank test, another non-parametric equivalent to the matched t.
  - `W:Normal` A test for normality. As you might infer from `Pr<W`, the associated p-value*lower* tail area of some distribution. If $p < 0.05$, conclude that the data are not normally distributed.

    The assumptions of the hypothesis tests for multiple regression imply that the residuals are normally distributed, though not quite independent. The lack of independence makes the W test a bit too likely to indicate lack of normality. If the test is non-significant, can one conclude that the data *are* normal? This is an example of a more general question: When can one conclude that the null hypothesis is true?

    To answer this question "Never" is just plain stupid, but still I don't want to go there right now. Instead, just two comments:

    * Like most tests, the W test for normality is much more sensitive when the sample size is large. So failure to observe a significant departure from normality does not imply that the data really are normal, for a small sample like this one (n=21).
    * In an observational study, residuals can appear non-normal because important independent variables have been omitted from the full model.
  - `Extremes` are the 5 highest and 5 lowest scores. Very useful for locating outliers. The largest residual in this data set is 20.21507; it's observation 5.
  - `Normal Probability Plot` is supposed to be straight-line if the data are normal. Even though I requested pagesize=35, this plot is pretty squashed. Basically it's useless.

- `proc plot` Does not show much of anything in this case. This is basically good news, though again the data are artificial. The default plotting symbol is `A`; if two points get too close together, they are plotted as `B`, and so on.

Here are a few sample questions.

Sample Question 3.5.13 *What is the mean of the average household incomes of the 21 towns?*

Answer to Sample Question 3.5.13 *$17,143*

Sample Question 3.5.14 *Is this the same as the average income of all the households in the 21 towns?*

Answer to Sample Question 3.5.14 *No way.*

Sample Question 3.5.15 *The custom test labelled* `MONEY` *is identical to what default test?*

Answer to Sample Question 3.5.15 *The t-test for* `INCOME`*. $F = t^2$, and the p-value is the same.*

Sample Question 3.5.16 *In the* `proc iml` *output, what can you learn from comparing* `F` *to* `FCRIT`*?*

Answer to Sample Question 3.5.16 $p < 0.05$

Sample Question 3.5.17 *For a town with 68,500 children 16 and under, and an average household income of $16,700, does the full model overpredict or underpredict sales? By how much?*

Answer to Sample Question 3.5.17 *Underpredict by $12,784. This is the first residual produced by* `proc print`*.*

Bibliography

[1] Cody, R. P. and Smith, J. K. (1991). *Applied statistics and the SAS programming language.* (4th Edition) Upper Saddle River, New Jersey: Prentice-Hall.

[2] Moore, D. S. and McCabe, G. P. (1993). *Introduction to the practice of statistics.* New York: W. H. Freeman.

[3] Neter, J., Kutner, M. H., Nachhtscheim, C. J. and Wasserman, W. (1996) *Applied linear statistical models.* (4th Edition) Toronto: Irwin.