Chapter 3

More Than One Independent Variable at a time

The standard elementary tests typically involve one independent variable and one dependent variable. Now we will see why this can make them very misleading. The lesson you should take away from this discussion is that when important variables are ignored in a statistical analysis — particularly in an observational study — the result can be that we draw incorrect conclusions from the data. Potential confounding variables really need to be included in the analysis.

3.1 The chi-squared test of independence

In order to make sure the central example in this chapter is clear, it may be helpful to give a bit more background on the common Pearson chi-square test of independence. As stated earlier, the chi-square test of independence is for judging whether two categorical variables are related or not. It is based upon a *cross-tabulation*, or *joint frequency distribution* of the two variables. For example, suppose that in the statclass data, we are interested in the relationship between sex and apparent ethnic background. If the ratio of females to males depended upon ethnic background, this could reflect an interesting cultural difference in sex roles with respect to men and women going to university (or at least, taking Statistics classes). In `statmarks1.sas`, we did this test and obtained a chisquare statistic of 2.92 (df=2, $p = 0.2321$), which is not statistically significant. Now we'll do it just a bit differently to illustrate the details. First, here is the program `ethsex.sas`.

```
/* ethsex.sas */
%include 'statread.sas';
title2 'Sex by Ethnic';
proc freq;
     tables sex*ethnic / chisq norow nocol nopercent expected;
```

And here is the output.

The FREQ Procedure

Table of sex by ethnic

sex ethnic(Apparent ethnic background (ancestry))

Frequency| Expected	Chinese	European	Other	Total
Male	27	7	5	39
	25.79	9.4355	3.7742	
Female	14	8	1	23
	15.21	5.5645	2.2258	
Total	41	15	6	62

Statistics for Table of sex by ethnic

Statistic	DF	Value	Prob
Chi-Square	2	2.9208	0.2321
Likelihood Ratio Chi-Square	2	2.9956	0.2236
Mantel-Haenszel Chi-Square	1	0.0000	0.9949
Phi Coefficient		0.2170	
Contingency Coefficient		0.2121	
Cramer's V		0.2170	

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

Sample Size = 62

In each cell of the table, we have an observed frequency and an expected frequency. The expected frequency is the frequency one would expect by chance if the two variables were completely unrelated.[1] If the observed frequencies are different enough from the expected frequencies, one would tend to disbelieve the null hypothesis that the two variables are unrelated. But how should one measure the difference, and what is the meaning of

[1]The formula for the expected frequency in a given cell is (row total) × (column total)/(sample size). This follows from the definition of independent events given in introductory probability: the events A and B are independent if $P(A \cap B) = P(A)P(B)$. But this is too much detail, and we're not going there.

different "enough?"

The Pearson chi-square statistic (named after Karl Pearson, a famous racist, uh, I mean statistician) is defined by

$$\chi^2 = \sum_{\text{cells}} \frac{(f_o - f_e)^2}{f_e}, \tag{3.1}$$

where f_o refers to the observed frequence, f_e refers to expected frequency, and as indicated, the sum is over all the cells in the table.

If the two variables are really independent, then as the total sample size increases, the probability distribution of this statistic approaches a chisquare with degrees of freedom equal to (Number of rows - 1)×(Number of columns - 1). Again, this is an approximate, large-sample result, one that obtains exactly only in the limit as the sample size approaches infinity. A traditional "rule of thumb" is that the approximation is okay if no expected frequency is less than five. This is why SAS gave us a warning.

More recent research suggests that to avoid inflated Type I error (false significance at a rate greater than 0.05), all you need is for no expected frequency to be less than one. You can see from formula (3.1) why an expected frequency less than one would be a problem. Division by a number close to zero can yield a very large quantity even when the observer and expected frequencies are fairly close, and the so-called chisquare value will be seriously inflated.

Anyway, The p-value for the chisquare test is the upper tail area, the area under the chi-square curve beyond the observed value of the test statistic. In the example from the statclass data, the test was not significant and we conclude nothing.

3.2 The Berkeley Graduate Admissions data

Now we're going to look at another example, one that should surprise you. In the 1970's the University of California at Berkeley was accused of discriminating against women in graduate admissions. Data from a large number of applicants are available. The three variables we will consider are sex of the person applying for graduate study, department to which the person applied, and whether or not the person was admitted. First, we will look at the table of sex by admission.

Table of sex by admit

sex	admit		
Frequency|			
Row Pct |	No	Yes	Total
Male	1493	1198	2691
	55.48	44.52	
Female	1278	557	1835
	69.65	30.35	
Total	2771	1755	4526

The FREQ Procedure

Statistics for Table of sex by admit

Statistic	DF	Value	Prob
Chi-Square	1	92.2053	<.0001

It certainly looks suspicious. Roughly forty-five percent of the male applicants were admitted, compared to thirty percent of the female applicants. This difference in percentages (equivalent to the relationship between variables here) is highly significant; with $n = 4526$, the p-value is very close to zero.

3.3 Controlling for a variable by subdivision

However, things look different when we take into account the department to which the person applied. Think of a *three-dimensional* table in which the rows are sex, the columns are admission, and the third dimension (call it layers) is department. Such tables are easy to generate with SAS and other statistical packages.

The three-dimensional table is displayed by printing each layer on a separate page, along with test statistics (if requested) for each sub-table. This is equivalent to dividing the cases into sub-samples, and doing the chisquare test separately for each sub-sample. A useful way to talk about this is to say that that we are *controlling* for the third variable; that is, we are looking at the relationship between the other two variables with the third variable held constant. We will have more to say about controlling for collections of independent variables when we get to regression.

Here are the six sub-tables of sex by admit, one for each department, with a brief comment after each table. The SAS output is edited a bit to save paper.

Table 1 of sex by admit
Controlling for dept=A

sex	admit		
Frequency|			
Row Pct	No	Yes	Total
Male	313	512	825
	37.94	62.06	
Female	19	89	108
	17.59	82.41	
Total	332	601	933

Statistics for Table 1 of sex by admit
Controlling for dept=A

Statistic	DF	Value	Prob
Chi-Square	1	17.2480	<.0001

For department A, 62% of the male applicants were admitted, while 82% of the female applicants were admitted. That is, women were *more* likely to get in than men. This is a *reversal* of the relationship that is observed when the data for all departments are pooled!

Table 2 of sex by admit
Controlling for dept=B

sex	admit		
Frequency| Row Pct	No	Yes	Total
Male	207 36.96	353 63.04	560
Female	8 32.00	17 68.00	25
Total	215	370	585

Statistics for Table 2 of sex by admit
Controlling for dept=B

Statistic	DF	Value	Prob
Chi-Square	1	0.2537	0.6145

For department B, women were somewhat more likely to be admitted (another reversal), but it's not statistically significant.

Table 3 of sex by admit
Controlling for dept=C

sex admit

Frequency| Row Pct	No	Yes	Total
Male	205	120	325
	63.08	36.92	
Female	391	202	593
	65.94	34.06	
Total	596	322	918

Statistics for Table 3 of sex by admit
Controlling for dept=C

Statistic	DF	Value	Prob
Chi-Square	1	0.7535	0.3854

For department C, men were slightly more likely to be admitted, but the 3% difference is much smaller than for the pooled data. Again, it's not statistically significant.

Table 4 of sex by admit
Controlling for dept=D

sex admit

Frequency| Row Pct	No	Yes	Total
Male	279	138	417
	66.91	33.09	
Female	244	131	375
	65.07	34.93	
Total	523	269	792

Statistics for Table 4 of sex by admit
Controlling for dept=D

Statistic	DF	Value	Prob
Chi-Square	1	0.2980	0.5852

For department D, women were a bit more likely to be admitted (a reversal), but it's far from statistically significant. Now department E:

Table 5 of sex by admit
Controlling for dept=E

sex admit

Frequency| Row Pct	No	Yes	Total
Male	138	53	191
	72.25	27.75	
Female	299	94	393
	76.08	23.92	
Total	437	147	584

Statistics for Table 5 of sex by admit
Controlling for dept=E

Statistic	DF	Value	Prob
Chi-Square	1	1.0011	0.3171

This time it's a non-significant tendency for men to get in more. Finally, department F:

Table 6 of sex by admit
Controlling for dept=F

sex admit

Frequency| Row Pct	No	Yes	Total
Male	351	22	373
	94.10	5.90	
Female	317	24	341
	92.96	7.04	
Total	668	46	714

Statistic	DF	Value	Prob
Chi-Square	1	0.3841	0.5354

For department F, women were slightly more likely to get in, but once again it's not significant.

So in summary, the pooled data show that men were more likely to be admitted to graduate study. But when take into account the department to which the student is applying, there is a significant relationship between sex and admission for only one department, and in that department, women are more likely to be accepted.

How could this happen? I generated two-way tables of sex by department and department by admit; both relationships were highly significant. Instead of displaying the SAS output, I have assembled some numbers from these two tables. The same thing could be accomplished with SAS `proc tabulate`, but it's too much trouble, so I did it by hand.

Table 3.1: Percentage of female applicants and overall percentage of applicants accepted for six departments

Department	Percent applicants female	Percentage applicants accepted
A	11.58%	64.42%
B	4.27	63.25
C	64.60	35.08
D	47.35	33.96
E	67.29	25.17
F	47.76	6.44

Now it is clear. The two departments with the lowest percentages of female applicants (A and B) also had the highest overall percentage of applicants accepted, while the department with the highest percentage of female applicants (E) also had the second-lowest overall percentage of applicants accepted. That is, the departments most popular with men were easiest to get into, and those most popular with women were more difficult. Clearly, this produced the overall tendency for men to be admitted more than women.

By the way, does this mean that the University of California at Berkeley was *not* discriminating against women? By no means. Why does a department admit very few applicants relative to the number who apply? Because they do not have enough professors and other resources to offer more classes. This implies that the departments popular with men were getting more resources, relative to the level of interest measured by number of applicants. Why? Maybe because men were running the show. The "show," by the way definitely includes the U. S. military, which funds a lot of engineering and similar stuff at big American universities.

The Berkeley data, a classic example of *Simpson's paradox*, illustrate the following uncomfortable fact about observational studies. When you include a new variable in an analysis, the results you have could get weaker, they could get stronger, or they could reverse direction — all depending upon the inter-relations of the independent variables. Basically, if an observational study does not include every potential confounding variable you can think of, there is going to be trouble.

Now, the distinguishing feature of the "elementary" tests is that they all involve one independent variable and one dependent variable. Consequently, they can be *extremely* misleading when applied to the data from observational studies, and are best used as tools for preliminary exploration.

Pooling the chi-square tests When using sub-tables to control for a categorical independent variable, it is helpful to have a single test that allows you to answer a question like this: If you control for variable A, is B related to C? For the chi-square test of independence, it's quite easy. Under the null hypothesis that B is unrelated to C for each value of A, the test statistics for the sub-tables are independent chisquare random variables. Therefore, there sum is also chisquare, with degrees of freedom equal to the sum of degrees of freedom for the sub-tables.

In the Berkeley example, we have a pooled chisquare value of

$$17.2480 + 0.2537 + 0.7535 + 0.2980 + 1.0011 + 0.3841 = 19.9384$$

with 6 degrees of freedom. Using any statistics text (except this one), we can look up the critical value at the 0.05 significance level. It's 12.59; since 19.9 ¿ 12.59, the pooled test is significant at the 0.05 level. To get a p-value for our pooled chisquare test, we can use SAS. See the program in the next section.

In summary, we need to use statistical methods that incorporate more than one independent variable at the same time; multiple regression is the central example. But even with advanced statistical tools, the most important thing in any study is to collect the right data in the first place. Looking at it the right way is critical too, but no statistical analysis can compensate for having the wrong data.

For more detail on the Berkeley data, see the article in *Science* by Bickel Hammel and O'Connell [1]. For the principle of adding chisquare values and adding degrees of freedom from sub-tables, a good reference is Feinberg's *The analysis of cross-classified categorical data* [3].

3.4 The SAS program

Here is the program `berkeley.sas`. It has several features that you have not seen yet, so a discussion follows the listing of the program.

```
/************************** berkeley.sas *********************************/
options linesize=79 pagesize=35 noovp formdlim='_';
title 'Berkeley Graduate Admissions Data: ';

proc format;
     value sexfmt 1 = 'Female' 0 = 'Male';
     value ynfmt 1 = 'Yes'  0 = 'No';
data berkley;
     input  line sex dept $ admit count;       %$
     format sex sexfmt.; format admit ynfmt.;
     datalines;
   1     0      A      1    512
   2     0      B      1    353
   3     0      C      1    120
   4     0      D      1    138
   5     0      E      1     53
   6     0      F      1     22
```

```
   7     1      A      1     89
   8     1      B      1     17
   9     1      C      1    202
  10     1      D      1    131
  11     1      E      1     94
  12     1      F      1     24
  13     0      A      0    313
  14     0      B      0    207
  15     0      C      0    205
  16     0      D      0    279
  17     0      E      0    138
  18     0      F      0    351
  19     1      A      0     19
  20     1      B      0      8
  21     1      C      0    391
  22     1      D      0    244
  23     1      E      0    299
  24     1      F      0    317
;
proc freq;
     tables sex*admit / nopercent nocol chisq;
     tables dept*sex / nopercent nocol chisq;
     tables dept*admit / nopercent nocol chisq;
     tables dept*sex*admit / nopercent nocol chisq;
     weight count;

/* Get p-value */
proc iml;
     x = 19.9384;
     pval = 1-probchi(x,6);
     print "Chisquare = " x "df=6, p = " pval;
```

The first unusual feature of `berkeley.sas` is in spite of recommendations to the contrary, the data are in the program itself rather than in a separate file. The data are in the data step, following the `datalines` command and ending with a semicolon. You can always do this, but usually it's a bad idea. Here, it's a good idea. This is why.

I did not have access to a raw data file, just a 2 by 6 by 2 table of sex by department by admission. So I just created a data set with 24 lines, even though there are 4526 cases. Each line of the data set has values for the three variables, and also a variable called `count`, which is just the observed cell frequency for that combination of sex, department and admission. Then, using the `weight` statement in `proc freq`, I just "weighted" each of the 24 cases in the data file by `count`, essentially multiplying the sample size by count for each case.

The advantages are several. First, such a data set is easy to create from published tables, and is much less trouble than a raw data file with thousands of cases. Second, the data file is so short that it makes sense to put it in the data set for portability and ease of reference. Finally, this is the way you can get the data from published tables (which may

not include any significance tests at all) into SAS, where you can compute any statistics you want, including sophisticated log-linear modelling analyses.

The last `tables` statement in the `proc freq` gives us the three-dimensional table. For a two-dimensional table, the first variable you mention will correspond to rows and the second will correspond to columns. For higher-dimensional tables, the second-to-last variable mentioned is rows, the last is columns, and combinations of the variables listed first are the control variables for which sub-tables are produced.

Finally, the `iml` in `proc iml` stands for "Interactive Matrix Language," and you can use it to perform useful calculations in a syntax that is very similar to standard matrix algebra notation; this can be very convenient when formulas you want to compute are in that notation. Here, we're just using it to calculate the area under the curve of the chisquare density with 6 degrees of freedom, beyond the observed test statistic of 19.9384. The `probchi` function is the cumulative distribution function of the chisquare distribution; the second argument (6 in this case) is the degrees of freedom. `probchi(`x`,6)` gives the area under the curve between zero and x, and `1-probchi(`x`,6)` gives the tail area above x – that is, the p-value.

Summary The example of the Berkeley graduate admissions data teaches us that potential confounding variables need to be explicitly included in a statistical analysis. Otherwise, the results can be very misleading. In the Berkeley example, first we ignored department and there was a relationship between sex and admission that was statistically significant in one direction. Then, when we *controlled* for department — that is, when we took it into account — the relationship was either significant in the opposite direction, or it was not significant (depending on which department).

We also saw how to pool chi-square values and degrees of freedom by adding over sub-tables, obtaining a useful test of whether two categorical variables are related, while controlling for one or more other categorical variables. This is something SAS will not do for you, but it's easy to do with `proc freq` output and a calculator.

Chapter 4

Multiple Regression: Part One

4.1 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. The analysis of the Berkeley graduate admissions data in Chapter 3 is our prime example. As another example where the relationship of interest is between quantitative rather than categorical variables, the correlation of education with 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.

4.2 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

$$E[Y|X = x] = \beta_0 + \beta_1 x. \tag{4.1}$$

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:

$$\widehat{Y} = b_0 + b_1 x. \tag{4.2}$$

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 (4.1) means that you can use formula (4.2) 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 4 *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 [5]. 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

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

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

$$= \beta_3$$

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 $w = a_1 z_1 + a_2 z_2 + a_3 z_3$, 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.

4.3 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.

4.3.1 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.

4.3.2 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 4.3.1 *Why is* $E[Y|X = x] = \beta_0 + \beta_1 x$ *(with* x *= Group) a silly model?*

Answer to Sample Question 4.3.1 *Designation of the Groups as 1, 2 and 3 is completely arbitrary.*

Sample Question 4.3.2 *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 4.3.2

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 4.3.3 *What does each* β *value mean?*

Answer to Sample Question 4.3.3 $\beta_0 = \mu_3$*, 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 4.3.4 *Why would it be nice to simultaneously test whether* β_1 *and* β_2 *are different from zero?*

Answer to Sample Question 4.3.4 *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 4.3.5 *What would happen if you used k indicator dummy variables instead of $k-1$?*

Answer to Sample Question 4.3.5 *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?

4.3.3 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.

4.4 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.

4.4.1 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 4.4.1 *Why is $SSR_F \geq SSR_R$?*

Answer to Sample Question 4.4.1 *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

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}, \tag{4.3}$$

where MSE_F is the mean square error for the full model: $MSE_F = \frac{SSE_F}{n-p}$. Equation 4.3 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 lecture. 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 4.4.2 *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 4.4.2

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 4.4.3 *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 4.4.3

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 4.4.4 *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 4.4.4

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 4.4.5 *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 4.4.5

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 4.4.6 *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 4.4.6

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 4.3 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 (4.3) 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 (4.3) 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 (4.3), 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.

4.4.2 Connections between Explained Variation and Significance Testing

If you divide numerator and denominator of Equation (4.3) 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 (4.3), we see that the relationship between F and a is

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right). \tag{4.4}$$

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

$$a = \frac{sF}{n - p + sF} \tag{4.5}$$

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 (4.5) 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.

4.5 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 [5]. 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 4.5.1 *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 4.5.1 $F = 99.103$, $p < 0.0001$

Sample Question 4.5.2 *What can you conclude from just this one test?*

Answer to Sample Question 4.5.2 *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 4.5.3 *What percent of the variation in sales is explained by number of kids 16 and under and average family income?*

Answer to Sample Question 4.5.3 *91.67%*

Sample Question 4.5.4 *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 4.5.4

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

Sample Question 4.5.5 *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 4.5.5

1. $t = 2.305$

2. $p = 0.0333$

3. Yes.

4. Positive.

Sample Question 4.5.6 *What do you conclude from this entire analysis? Direct your answer to a statistician or researcher.*

Answer to Sample Question 4.5.6 *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 4.5.7 *What do you conclude from this entire analysis? Direct your answer to a person without statistical training.*

Answer to Sample Question 4.5.7 *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 4.5.8 *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 4.5.8 $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2$ *= -68.857073 + 1.454560*65.4 + 9.365500*17.6 = 191.104, so predicted annual sales = $191,104.*

Sample Question 4.5.9 *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 4.5.9 *Predicted annual sales goes up by $1,454.*

Sample Question 4.5.10 *What do you conclude from the t-test for the intercept?*

Answer to Sample Question 4.5.10 *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 4.5.11 *Controlling for average household income, what proportion of the remaining variation is explained by number of children under 16?*

Answer to Sample Question 4.5.11 *Using* $F = t^2$ *and plugging into (4.5), 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 4.5.12 *Controlling for number of children under 16, what proportion of the remaining variation is explained by average household income?*

Answer to Sample Question 4.5.12 $a = \frac{2.305^2}{18+2.305^2} = 0.2278994$*, 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 4.3 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 (4.5) 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* [2]. 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
       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

                             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


     Variable  DF     Type I SS

     INTERCEP   1        694876
     KIDS       1         23372
     INCOME     1    643.475809
        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al
3
                           With bells and whistles
                                            10:58 Saturday, January 19, 2002

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

        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al
4
                           With bells and whistles
                                            10:58 Saturday, January 19, 2002

                                    FCRIT
                                 4.4138734


                                       F       TSQ      PVAL
                                5.3108439  5.313025 0.0333214


                                                                      A
                     Proportion of remaining variation is  0.2278428

        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al
```

```
5
                          With bells and whistles
                                        10:58 Saturday, January 19, 2002

                        General Linear Models Procedure

                    Number of observations in data set = 21

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

                        General Linear Models Procedure

Dependent Variable: SALES
                                    Sum of            Mean
Source                    DF          Squares          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
        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al
7
                          With bells and whistles
                                        10:58 Saturday, January 19, 2002

                        General Linear Models Procedure

Dependent Variable: SALES

                                     T for H0:    Pr > |T|   Std Error of
Parameter                  Estimate    Parameter=0                Estimate

Xh p249                  191.103930         69.07     0.0001     2.76679783


                                     T for H0:    Pr > |T|   Std Error of
Parameter                  Estimate    Parameter=0                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


        Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      8
                          With bells and whistles
                                        11:32 Tuesday, January 15, 2002

             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

9

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		

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

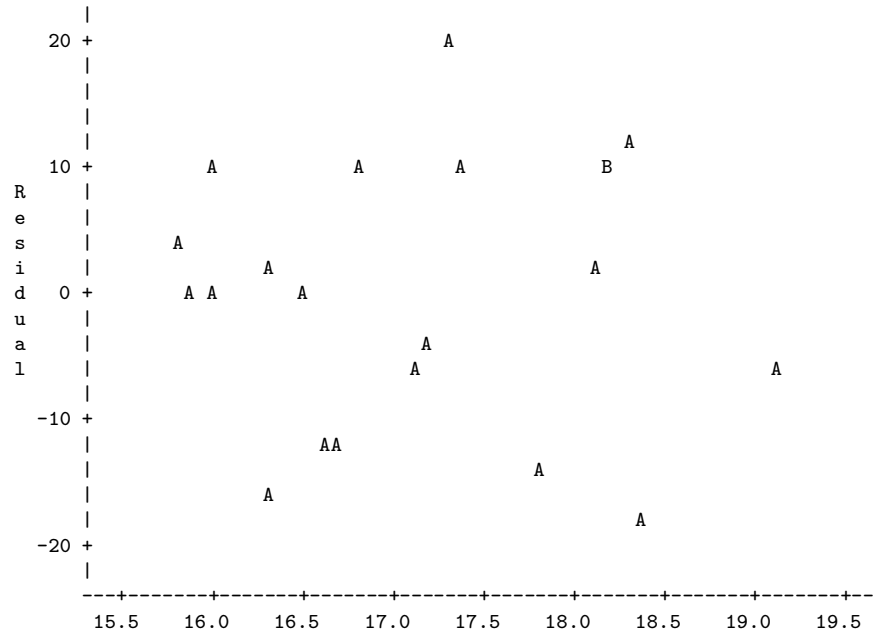
11:32 Tuesday, January 15, 2002

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

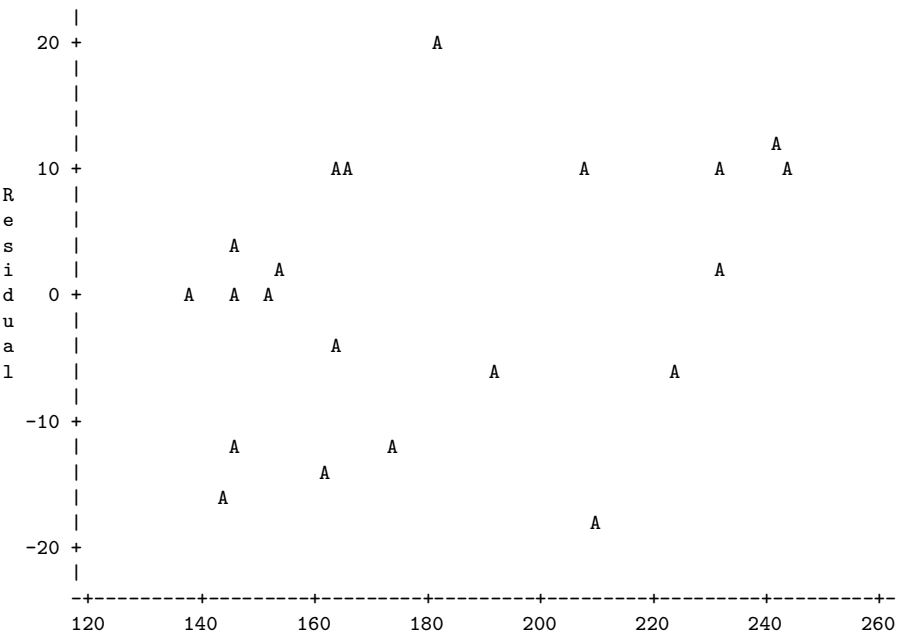
```
     |
  20 +                A
     |
     |
     |
     |                                                    A
  10 +              A      A             A                A  A
R    |
e    |
s    |            A
i    |                A                                     A
d   0 +        A  A        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 4.5.12.
- `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 4.5.8. 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 4.5.13 *What is the mean of the average household incomes of the 21 towns?*

Answer to Sample Question 4.5.13 *$17,143*

Sample Question 4.5.14 *Is this the same as the average income of all the households in the 21 towns?*

Answer to Sample Question 4.5.14 *No way.*

Sample Question 4.5.15 *The custom test labelled* `MONEY` *is identical to what default test?*

Answer to Sample Question 4.5.15 *The t-test for* `INCOME`*. $F = t^2$, and the p-value is the same.*

Sample Question 4.5.16 *In the* `proc iml` *output, what can you learn from comparing* `F` *to* `FCRIT`*?*

Answer to Sample Question 4.5.16 $p < 0.05$

Sample Question 4.5.17 *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 4.5.17 *Underpredict by $12,784. This is the first residual produced by* `proc print`*.*

Bibliography

[1] Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, **187**, 398-403.

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

[3] Feinberg, S. (1977) *The analysis of cross-classified categorical data.* Cambridge, Massachusetts: MIT Press.

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

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