Chapter 5

Multiple Regression: Part One

5.1 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 independent variable and dependent variable, 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. And a very good example is provided by the `tubes` data of Chapter 3, where experimental conditions were so tightly controlled that there was practically no available source of variation in growth rate except for the genetic character of the fungus.

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 4 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.

5.2 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

$$E[Y|X = x] = \beta_0 + \beta_1 x. \tag{5.1}$$

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:

$$\widehat{Y} = b_0 + b_1 x. \tag{5.2}$$

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 (5.1) means that you can use formula (5.2) to estimate the population mean of Y for a given value of $X = x$ *without having any data* at that x value. This is not "cheating," or at any rate, it need not be. If

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

then the estimate can be quite good.

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

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

Data Analysis Hint 3 *There is a direct tradeoff between amount of data and the strength (restrictiveness) of model assumptions. If you have a lot of data, you do not need to assume as much. If you have a small sample, 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 [16]. 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

$$\begin{array}{rlll} & \beta_0 + \beta_1 x_1 + \beta_2 x_2 & +\beta_3(x_3+1) & +\beta_4 x_4 \\ - & (\beta_0 + \beta_1 x_1 + \beta_2 x_2 & +\beta_3 x_3 & +\beta_4 x_4) \end{array} \tag{5.3}$$

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

$$= \beta_3$$

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 $\beta_k < 0$.

Recall from Chapter 3 that 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.

5.3 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.

5.3.1 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 independent variable?

- 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 dependent variable may depend on the independent variable 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.

5.3.2 Categorical Independent Variables

Indicator Dummy 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 5.3.1 *Why is $E[Y|X = x] = \beta_0 + \beta_1 x$ (with x = Group) a silly model?*

Answer to Sample Question 5.3.1 *Designation of the Groups as 1, 2 and 3 is completely arbitrary.*

Sample Question 5.3.2 *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 5.3.2

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 5.3.3 *What does each β value mean?*

Answer to Sample Question 5.3.3 *$\beta_0 = \mu_3$, 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 5.3.4 *Why would it be nice to simultaneously test whether β_1 and β_2 are different from zero?*

Answer to Sample Question 5.3.4 *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).*

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.

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.

Add a quantitative independent variable

Now suppose we include patient's age in the regression model. When there are both quantitative and categorical independent variables, the quantitative variables are often called *covariates*, particularly if the categorical part is experimentally manipulated. Tests of the categorical variables controlling for the quantitative variables are called *analysis of covariance*.

The usual practice is to put the covariates first. So, we'll let X_1 represent age, and let X_2 and X_3 be the indicator dummy variables for experimental condition. The model now is that all conditional distributions are normal with the same variance σ^2, and population mean

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

Sample Question 5.3.5 *Fill in the table.*

Group	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Answer to Sample Question 5.3.5

Group	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$\mu_2 = (\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\mu_3 = \beta_0 + \beta_1 x_1$

This is a *parallel slopes model*. That is, there is a least-squares regression line for each group, with the same slope β_1 for each line. Only the intercepts are different. This means that for any fixed value of x_1 (age), the differences among population means are the same. For any value of age (that is, holding age constant, or *controlling* for age), the difference between response to Drug A and the placebo is β_2. And controlling for age), the difference between response to Drug B and the placebo is β_3. The three group means are equal for each constant value of age if (and only if) $\beta_2 = \beta_3 = 0$. This is the null hypothesis for the analysis of covariance.

It is easy (and often very useful) to have more than one covariate. In this case we have parallel planes or hyper-planes. And at any fixed set of covariate values, the distances among hyperplanes correspond exactly to the differences among the intercepts. This means we are usually interested in testing null hypotheses about the regression coefficients corresponding to the dummy variables.

Sample Question 5.3.6 *Suppose we want to test the difference between response to Drug A and Drug B, controlling for age. What is the null hypothesis?*

Answer to Sample Question 5.3.6 $H_0 : \beta_2 = \beta_3$

Sample Question 5.3.7 *Suppose we want to test whether controlling for age, the average response to Drug A and Drug B is different from response to the placebo. What is the null hypothesis?*

Answer to Sample Question 5.3.7 $H_0 : \beta_2 + \beta_3 = 0$

Sample Question 5.3.8 *Huh? Show your work.*

Answer to Sample Question 5.3.8

$$\frac{1}{2}[\,(\beta_0+\beta_2+\beta_1x_1)+(\beta_0+\beta_3+\beta_1x_1)\,]=\beta_0+\beta_1x_1$$

$$\iff \quad \beta_0+\beta_2+\beta_1x_1+\beta_0+\beta_3+\beta_1x_1=2\beta_0+2\beta_1x_1$$

$$\iff \quad 2\beta_0+\beta_2+\beta_3+2\beta_1x_1=2\beta_0+2\beta_1x_1$$

$$\iff \quad \beta_2+\beta_3=0$$

The symbol $\iff$ means "if and only if." The arrows can logically be followed in both directions.

This last example illustrates several important points.

- Contrasts can be tested with indicator dummy variables.
- If there are covariates, the ability to test contrasts *controlling* for the covariates is very valuable.
- Sometimes, the null hypothesis for a contrast of interest might not be what you expect, and you might have to derive it algebraically. This can be inconvenient, and it is too easy to make mistakes.

Cell means coding

When students are setting up dummy variables for a categorical independent variable with p categories, the most common mistake is to define an indicator dummy variable for every category, resulting in p dummy variables rather than $p-1$ — and of course there is an intercept too, because it's a regression model and regression software almost always includes an intercept unless you explicitly suppress it. But then the p population means are represented by $p+1$ regression coefficients, and mathematically, the representation cannot be unique. In this situation the least-squares estimators are not unique either, and all sorts of technical problems arise. Your software might try to save you by throwing one of the dummy variables out, but which one would it discard? And would you notice that it was missing from your output?

Suppose, however, that you used p dummy variables but *no intercept* in the regression model. Then there are p regression coefficients corresponding to the p population means, and all the technical problems go away. The correspondence between regression coefficients and population means is unique, and the model can be handy. In particular, null hypotheses can often be written down immediately without any high school algebra. Here is how it would look for the study with two drugs and a placebo. The conditional population means is

$$E[Y|\boldsymbol{X}=\boldsymbol{x}]=\beta_1x_1+\beta_2x_2+\beta_3x_3,$$

and the table of population means has a very simple form:

Drug	x_1	x_2	x_3	$\beta_1x_1+\beta_2x_2+\beta_3x_3$
A	1	0	0	$\mu_1=\beta_1$
B	0	1	0	$\mu_2=\beta_2$
Placebo	0	0	1	$\mu_3=\beta_3$

The regression coefficients correspond directly to population (cell) means for any number of categories; this is why it's called *cell means coding*. Contrasts are equally easy to write in terms of μ or β quantities.

Cell means coding works nicely in conjunction with quantitative covariates. In the drug study example, represent age by X_4. Now the conditional population mean is

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

and the cell means (for any fixed value of age equal to x_4) are

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

This is another parallel slopes model, completely equivalent to the earlier one. The regression coefficients for the dummy variables are the intercepts, and because the lines are parallel, the differences among population means at any fixed value of x_4 are exactly the differences among intercepts. Note that

- It is easy to write the null hypothesis for any contrast of collection of contrasts. Little or no algebra is required.
- This extends to categorical independent variables with any number of categories.
- With more than one covariate, we have a parallel planes model, and it is still easy to express the null hypotheses.
- The `test` statement of `proc reg` is a particularly handy tool.

Effect Coding

In *effect coding* there are $p - 1$ dummy variables for a categorical independent variable with p categories, and the intercept is included. Effect coding look just like indicator dummy variable coding with an intercept, except that the last (reference) category gets -1 instead of zero. Here's how it looks for the hypothetical drug study.

Group	x_1	x_2	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

To see what the regression coefficients mean, first define μ to be the average of the three population means. Then

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0,$$

so that the intercept is the mean of population means — sometimes called the *grand mean*. Now we can see right away that

- β_1 is the difference between μ_1 and the grand mean.
- β_2 is the difference between μ_2 and the grand mean.
- $-\beta_1 - \beta_2$ is the difference between μ_3 and the grand mean.
- Equal population means is equivalent to zero coefficients for all the dummy variables.
- The last category is not a reference category. It's just the category with the least convenient expression for the deviation from the grand mean.
- This pattern holds for any number of categories.

In the standard language of analysis of variance, *effects* are deviations from the grand mean. That's why this dummy variable coding scheme is called "effect coding." When there is more than one categorical independent variable, the average cell mean for a particular category (averaging across other independent variables) is called a *marginal mean*, and the so-called *main effects* are deviations of the marginal means from the grand mean; these are represented nicely by effect coding. Equality of marginal means implies that all main effects for the variable are zero, and vice versa.

Sometimes, people speak of testing for the "main effect" of a categorical independent variable. This is a loose way of talking, because there is not just one main effect for a variable. There are at least two, one for each marginal mean. Possibly, this use of "effect" blends the effect of an experimental variable with the technical statistical meaning of effect. However, it's a way of talking that does no real harm, and you may see it from time to time in this text.

We will see later that effect coding is very useful when there is more than one categorical independent variable and we are interested in *interactions* — ways in which the relationship of an independent variable with the dependent variable depends on the value of another independent variable.

Covariates work nicely with effect coding. There is no need to make a table of expected values, unless a question explicitly asks you to do so. For example, suppose you add the covariate X_1 = Age to the drug study. The treatment means (which depend on X_1 are as follows:

Group	x_2	x_3	$E[Y\|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1). So, if the regression coefficients for all the dummy variables equal zero, the categorical independent variable is unrelated to the dependent variable, when one controls for the covariates.

Finally, it's natural for a student to wonder: What dummy variable coding scheme should I use? Use whichever is most convenient. They are all equivalent, if done correctly. They yield the same test statistics, and the same conclusions.

5.3.3 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.

5.4 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 independent variable 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.

5.4.1 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 5.4.1 *Why is $SSR_F \geq SSR_R$?*

Answer to Sample Question 5.4.1 *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

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}, \tag{5.4}$$

where MSE_F is the mean square error for the full model: $MSE_F = \frac{SSE_F}{n-p}$. Equation 5.4 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.

Looking at the Formula for F

Formula 5.4 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 (5.4) 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 (5.4) 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 (5.4), 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.

5.4.2 Connections between Explained Variation and Significance Testing

If you divide numerator and denominator of Equation (5.4) 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 $R_F^2 - R_R^2$, but on a quantity that will be denoted by

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}. \tag{5.5}$$

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 (5.4), we see that the relationship between F and a is

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right). \tag{5.6}$$

so that for any given sample size, the bigger a is, the bigger F becomes. 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 (5.6) for F can be turned around to express a in terms of F, as follows:

$$a = \frac{sF}{n - p + sF} \tag{5.7}$$

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 (5.7) 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.

5.5 Measurement error

In a survey, suppose that a respondent's annual income is "measured" by simply asking how much he or she earned last year. Will this measurement be completely accurate? Of course not. Some people will lie, some will forget and give a reasonable guess, and still others will suffer from legitimate confusion about what constitutes income. Even physical variables like height, weight and blood pressure are subject to some inexactness of measurement, no matter how skilled the personnel doing the measuring. Many categorical variables are subject to *classification error*; a case is recorded as being in one category, but the truth is that it's in another. In fact, very few of the variables in the typical data set are measured completely without error.

So, there are really two versions of most variables – the *true* version and the *observed* version. Typically, the relationships we are interested in are relationships among the true variables, while the statistical analysis is necessarily based upon what we can observe.

In general, when there is a relationship between two true variables, the relationship also appears between the observed variables, but it is weaker. This means that things are not so bad when we are just testing for association between pairs of variables, and not trying to control for anything. But when we test for a relationship controlling for some set of variables, we are seeking it in the conditional distributions — that is, in the joint distributions of the independent and dependent variables, conditional on the values of the variables for which we are controlling. The unfortunate truth is this. If the control variables are measured with error, the conditional relationship given the observed variables need not be the same as the conditional relationship given the true variables.

It's as if we are trying to hold the control variables steady, but we can't tell exactly where they are. So the holding constant does not quite work. This applies to the model-based control of the classical regression models, and also to control by subdivision (if there is classification error in the categorical control variables). It even applies to experimental control, if it is not done *very* carefully.[1]

For example, suppose the subjects in a study are adults, and you are testing the relationship of age to Body Mass Index (BMI)[2], controlling for exercise and calorie intake. The questionnaire measures of exercise and calorie intake are known to be inaccurate. People exaggerate amount of exercise and under-report calories — and not by a constant amount. You can't see these control variables clearly to hold them constant. The result is that even if age is unrelated to Body Mass Index for every combination of *true* exercise and calorie intake, a relationship between age and BMI can exist conditionally upon *observed* exercise and calorie intake.

[1] Suppose a drug is being injected into a rat. The amount of drug injected may be exactly the same for all the rats in a particular experimental condition, but because of microscopic variation in needle placement and the rats' circulatory systems, the amount of drug that actually gets into the blood (the true dosage) may vary quite a bit. I am grateful to Prof. Alison Fleming for this remark.

[2] Weight in kilograms divided by the square of height in meters. Values above 25 are supposed to indicate obesity.

The poison combination Here is the situation that causes multiple regression to fail. You want to test B controlling for A.

1. A is related to the dependent variable
2. A and B are related to each other, and
3. A is measured with error.[3]

In this situation it is very tempting (and common practice) to just use the imperfect version of A, and try controlling for it with ordinary least-squares regression. But if you do this, all hell breaks loose. The regression coefficients b are biased estimators of the true regression coefficients β. Furthermore, the Type I Error can be badly inflated. In a 2009 paper, Brunner and Austin [3] point out that the problem biased estimation has been known since the 1930s. They also show that even for small amounts of measurement error and moderate sample sizes, the probability of falsely rejecting the null hypothesis at the 0.05 level can be unacceptably large. As the sample size increases, the probability of false significance approaches 1.00.

The problem of measurement error can be particularly acute in observational medical research. There, a common goal is to assess potential risk factors, controlling for known risk factors. The known risk factors (set A) do matter, and they are generally correlated with the potential risk factors that are being investigated. Also, the known risk factors are difficult to measure without error. In this situation, application of standard methods will often lead to the conclusion that the potential risk factors (set B) are a problem even when one controls for the known risk factors. Such conclusions are very suspect.

But all is not lost The problem really comes from trying to use regression as a *causal* model for observational data. As long as you are interested in prediction rather than interpretation, there is no problem. The test for whether age is a useful predictor of Body Mass Index is still valid, even if its usefulness comes from its correlation with true (as opposed to reported) exercise level. Viewed this way, the measurement error problem is just another reason why most observational studies cannot provide good evidence that the independent variable *causes*, or *contributes to* the dependent variable.

Also, of one or more categorical independent variables are experimentally manipulated, analysis of covariance can help reduce MSE and makes the analysis more precise, even if the covariates (control variables) are measured with error. There is no inflation of Type I Error rate because random assignment breaks up any association between A and B.

Even for purely observational studies, there are statistical methods that incorporate measurement error into the model, and are not subject to the bias and Type I error rate problems described here. One example is the structural equation models [2, 14]

[3]Measurement error in B, the variable of set of variables you are testing, does not matter much. In fact, it makes the problems described here a little less severe. This is also true of measurement error in the dependent variable. All this assumes that the errors of measurement are uncorrelated with each other. Correlated measurement error, which often arises from sloppy research design, introduces a set of problems that are usually fatal to correct inference.

popular in the social and (to a lesser degree) in the biological sciences. There are also more mainstream statistical methods [4, 10, 12] that overlap with structural equation modeling and have many of the same objectives. But all of these techniques require additional information (more variables, not just more cases) so that measurement error can be assessed. For example, it is helpful to have two independent readings of each independent variable. The main point is that it's never just a question of running different software. Studies need to be *planned* differently if measurement error is to be taken into account.

Rules of discourse A later edition of this text will have a chapter on structural equation models. In the meantime, we will apply traditional methods with care, and we will be careful how we talk about things. In any academic discipline, a great deal of the discussion tends to be repetitive, with the participants going back and forth, making points and counter-points that are well rehearsed, and very similar to what they have read or heard somewhere. The details may be about the case at hand, but the overall pattern is very predictable if you know the field. Think of these patterns of discussion as pathways, worn into the surface of scientific debate.[4] One thing that happens in a university education is that you are exposed to a fairly large collection of such scripts, and if you learn to apply them flexibly, it may help you think your way through difficult problems. If not, at least you sound educated.

Anyway, here are some ways to criticize an observational study based on measurement error in the independent variables.

- Start with "How did you control for . . . ?" The answer will be some version of subdivision or model-based control.
- Most of the time, it will be obvious that the control variables are measured with error. If so, ask "How did you take measurement error into account?" You might get a straight answer.
- If not, ask if they are aware of how ignoring measurement error in the independent variables can result in biased estimation and inflated Type I error rate. Provide references (given above).
- If they say "Oh, there was just a little measurement error," observe that if the sample is large enough, no amount of measurement error is safe. Brunner and Austin (2009) give a proof.
- If they say "Well, its the best we could do," you could ask whether its better to say something incorrect, or to be silent.

In this course, we will carry out classical regression analysis on observational data only when our primary purpose is prediction, and we will be very careful about the way we describe the results. We will use regression methods extensively on experimental data.

[4] One could make a similar point about political discussion, or even conversations about hockey.

5.6 Multiple Regression with SAS

It is often 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 [16]. The cases 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 noovp formdlim='_';
title 'Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al';
title2 'Just the defaults';

data portrait;
     infile 'dwaine.data';
     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

                              The REG Procedure
                                Model: MODEL1
                           Dependent Variable: sales

                     Number of Observations Read          21
                     Number of Observations Used          21


                              Analysis of Variance

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

Model                       2            24015          12008      99.10    <.0001
Error                      18       2180.92741      121.16263
Corrected Total            20            26196


               Root MSE             11.00739    R-Square     0.9167
               Dependent Mean      181.90476    Adj R-Sq     0.9075
               Coeff Var             6.05118


                              Parameter Estimates

                           Parameter       Standard
     Variable      DF       Estimate          Error    t Value    Pr > |t|

     Intercept      1      -68.85707       60.01695      -1.15      0.2663
     kids           1        1.45456        0.21178       6.87      <.0001
     income         1        9.36550        4.06396       2.30      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.
- `Root MSE` is the square root of Mean Square Error (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 independent variables.

- `Parameter Estimates` are the b values corresponding to the independent variables listed. The one corresponding to `Intercept` is b_0. `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, testing whether the regression coefficient is zero.

Here are some sample questions based on the list file.

Sample Question 5.6.1 *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 5.6.1 $F = 99.103$, $p < 0.0001$

Sample Question 5.6.2 *What can you conclude from just this one test?*

Answer to Sample Question 5.6.2 *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 5.6.3 *What percent of the variation in sales is explained by number of kids 16 and under and average family income?*

Answer to Sample Question 5.6.3 *91.67%*

Sample Question 5.6.4 *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 5.6.4

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

Sample Question 5.6.5 *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 5.6.5

1. $t = 2.305$
2. $p = 0.0333$
3. Yes.
4. Positive.

Sample Question 5.6.6 *What do you conclude from this entire analysis? Direct your answer to a statistician or researcher.*

Answer to Sample Question 5.6.6 *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 5.6.7 *What do you conclude from this entire analysis? Direct your answer to a person without statistical training.*

Answer to Sample Question 5.6.7 *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 5.6.8 *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 5.6.8 $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2$ = *-68.857073 + 1.454560*65.4 + 9.365500*17.6 = 191.104, so predicted annual sales = $191,104.*

Sample Question 5.6.9 *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 5.6.9 *Predicted annual sales goes up by $1,454.*

Sample Question 5.6.10 *What do you conclude from the t-test for the intercept?*

Answer to Sample Question 5.6.10 *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. In the published literature, sometimes all you have are reports of t-tests for regression coefficients.

Sample Question 5.6.11 *Controlling for average household income, what proportion of the remaining variation is explained by number of children under 16?*

Answer to Sample Question 5.6.11 *Using $F = t^2$ and plugging into (5.7), 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 5.6.12 *Controlling for number of children under 16, what proportion of the remaining variation is explained by average household income?*

Answer to Sample Question 5.6.12 $a = \frac{2.305^2}{18+2.305^2} = 0.2278994$*, 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 `proc reg` can be.

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

data portrait;
     infile 'dwaine.data';
     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 with the `ss1` option if you want to see them. The independent variables in the `model` statement are added to the model in order. For each variable, the `Type I SS` is 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. The t-tests correspond to `proc glm`'s Type III sums of squares; everything is controlled for everything else.

- `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 5.4, and obtain its p-value. This F should be equal to the square of the t statistic

from the printout, so we check. Then we use (5.7) 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 compared to 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* [5]. 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

                              The REG Procedure

                    Number of Observations Read          21
                    Number of Observations Used          21


                           Descriptive Statistics

                                          Uncorrected                    Standard
Variable            Sum          Mean            SS      Variance     Deviation

Intercept       21.00000       1.00000       21.00000             0             0
kids          1302.40000      62.01905           87708     346.71662      18.62033
income         360.00000      17.14286      6190.26000       0.94157       0.97035
sales         3820.00000     181.90476          721072    1309.81048      36.19130
```

Correlation

Variable	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

The REG Procedure
Model: MODEL1
Dependent Variable: sales

Number of Observations Read	21
Number of Observations Used	21

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24015	12008	99.10	<.0001
Error	18	2180.92741	121.16263		
Corrected Total	20	26196			

Root MSE	11.00739	R-Square	0.9167
Dependent Mean	181.90476	Adj R-Sq	0.9075
Coeff Var	6.05118		

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 3
With bells and whistles

The REG Procedure
Model: MODEL1
Dependent Variable: sales

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS
Intercept	1	-68.85707	60.01695	-1.15	0.2663	694876
kids	1	1.45456	0.21178	6.87	<.0001	23372
income	1	9.36550	4.06396	2.30	0.0333	643.47581

```
Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      4
                     With bells and whistles

                        The REG Procedure
                          Model: MODEL1

         Test allivs Results for Dependent Variable sales

                                    Mean
     Source             DF         Square    F Value    Pr > F

     Numerator           2          12008      99.10    <.0001
     Denominator        18      121.16263


Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      5
                     With bells and whistles

                        The REG Procedure
                          Model: MODEL1

          Test inter Results for Dependent Variable sales

                                    Mean
     Source             DF         Square    F Value    Pr > F

     Numerator           1       159.48430       1.32    0.2663
     Denominator        18      121.16263


Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      6
                     With bells and whistles

                        The REG Procedure
                          Model: MODEL1

          Test child Results for Dependent Variable sales

                                    Mean
     Source             DF         Square    F Value    Pr > F

     Numerator           1     5715.50583      47.17    <.0001
     Denominator        18      121.16263


Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      7
                     With bells and whistles

                        The REG Procedure
                          Model: MODEL1

          Test money Results for Dependent Variable sales

                                    Mean
     Source             DF         Square    F Value    Pr > F

     Numerator           1       643.47581       5.31    0.0333
     Denominator        18      121.16263
```

```
      Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al      8
                           With bells and whistles

                                    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      9
                           With bells and whistles

                              The GLM Procedure

                     Number of Observations Read          21
                     Number of Observations Used          21


      Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al     10
                           With bells and whistles

                              The GLM Procedure

Dependent Variable: sales

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

Model                         2    24015.28211    12007.64106     99.10   <.0001

Error                        18     2180.92741      121.16263

Corrected Total              20    26196.20952

                R-Square     Coeff Var      Root MSE    sales Mean

                0.916746      6.051183      11.00739      181.9048

Source                       DF      Type I SS    Mean Square   F Value   Pr > F

kids                          1    23371.80630    23371.80630    192.90   <.0001
income                        1      643.47581      643.47581      5.31   0.0333


Source                       DF    Type III SS    Mean Square   F Value   Pr > F

kids                          1    5715.505835    5715.505835     47.17   <.0001
income                        1     643.475809     643.475809      5.31   0.0333
```

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 11
With bells and whistles

The GLM Procedure

Dependent Variable: sales

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Xh p249	191.103930	2.76679783	69.07	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-68.85707315	60.01695322	-1.15	0.2663
kids	1.45455958	0.21178175	6.87	<.0001
income	9.36550038	4.06395814	2.30	0.0333

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 12
With bells and whistles

Obs	kids	income	sales	presale	resale
1	68.5	16.7	174.4	187.184	-12.7841
2	45.2	16.8	164.4	154.229	10.1706
3	91.3	18.2	244.2	234.396	9.8037
4	47.8	16.3	154.6	153.329	1.2715
5	46.9	17.3	181.6	161.385	20.2151
6	66.1	18.2	207.5	197.741	9.7586
7	49.5	15.9	152.8	152.055	0.7449
8	52.0	17.2	163.2	167.867	-4.6666
9	48.9	16.6	145.4	157.738	-12.3382
10	38.4	16.0	137.2	136.846	0.3540
11	87.9	18.3	241.9	230.387	11.5126
12	72.8	17.1	191.1	197.185	-6.0849
13	88.4	17.4	232.0	222.686	9.3143
14	42.9	15.8	145.3	141.518	3.7816
15	52.5	17.8	161.1	174.213	-13.1132
16	85.7	18.4	209.7	228.124	-18.4239
17	41.3	16.5	146.4	145.747	0.6530
18	51.7	16.3	144.0	159.001	-15.0013
19	89.6	18.1	232.6	230.987	1.6130
20	82.7	19.1	224.1	230.316	-6.2161
21	52.3	16.0	166.5	157.064	9.4356

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 13
With bells and whistles

The UNIVARIATE Procedure
Variable: resale (Residual)

Moments

N	21	Sum Weights	21
Mean	0	Sum Observations	0
Std Deviation	10.442527	Variance	109.046371
Skewness	-0.0970495	Kurtosis	-0.7942686
Uncorrected SS	2180.92741	Corrected SS	2180.92741
Coeff Variation	.	Std Error Mean	2.27874622

Basic Statistical Measures

Location		Variability	
Mean	0.000000	Std Deviation	10.44253
Median	0.744918	Variance	109.04637
Mode	.	Range	38.63896
		Interquartile Range	15.65166

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	2.5	Pr >= |M|	0.3833
Signed Rank	S	1.5	Pr >= |S|	0.9599

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 14
With bells and whistles

The UNIVARIATE Procedure
Variable: resale (Residual)

Tests for Normality

Test	--Statistic---		-----p Value------	
Shapiro-Wilk	W	0.954073	Pr < W	0.4056
Kolmogorov-Smirnov	D	0.147126	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.066901	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.432299	Pr > A-Sq	>0.2500

Quantiles (Definition 5)

Quantile	Estimate
100% Max	20.215072
99%	20.215072
95%	11.512629
90%	10.170574
75% Q3	9.435601
50% Median	0.744918
25% Q1	-6.216062
10%	-13.113212
5%	-15.001313
1%	-18.423890
0% Min	-18.423890

Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al 15
With bells and whistles

The UNIVARIATE Procedure
Variable: resale (Residual)

```
                Extreme Observations

      ------Lowest-----        ------Highest-----

         Value      Obs             Value      Obs

      -18.4239       16           9.75858        6
      -15.0013       18           9.80368        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
```

```
Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al     16
                    With bells and whistles

                  The UNIVARIATE Procedure
                 Variable:  resale  (Residual)

                       Normal Probability Plot
       22.5+                                          *****
           |                                       +++++
           |                                  +++++
           |                              ******
           |                         *******
        2.5+                      ++++
           |                 +++++
           |           +++++ *
           |      ******
      -17.5+
            +----+----+----+----+----+----+----+----+----+----+
                -2        -1         0        +1        +2
```

```
      Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al     17
                          With bells and whistles

           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                                 A         A
l     |
      |
  -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.

      |
      |
   20 +                              A
      |
      |
      |
      |                                            A
   10 +          A            A        A            B
R     |
e     |
s     |       A
i     |               A                           A
d   0 +        A A       A
u     |
a     |                            A
l     |                           A                                 A
      |
  -10 +
      |                   AA
      |                                     A
      |               A
      |                                           A
  -20 +
      |
      ---+-------+-------+-------+-------+-------+-------+-------+-------+--
       15.5    16.0    16.5    17.0    17.5    18.0    18.5    19.0    19.5

                                    income
```

```
     Dwaine Studios Example from Chapter 6 (Section 6.9) of Neter et al     19
                          With bells and whistles

          Plot of resale*sales.  Legend: A = 1 obs, B = 2 obs, etc.

      |
      |
   20 +                               A
      |
      |
      |
      |                                                          A
   10 +                       AA                   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 +
      |
      -+---------+---------+---------+---------+---------+---------+---------+-
      120       140       160       180       200       220       240       260

                                     sales
```

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 5.6.12.

- `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 — except that the denominator of the F ratio is MSE from the model including *all* the independent variables.

  - 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 $\widehat{Y}$ value, of 191.103930, which is what we got with a calculator as the answer to Sample Question 5.6.8. 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`. `proc glm` produces these by default only when none of the independent variables is declared categorical with the `class` statement. If you have categorical independent variables, you can request parameter estimates with the `parms` option.

- `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 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 is the *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? This question was discussed a bit in Chapter 1. Here are two additional comments about the tests for normality:
    - 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 5.6.13 *What is the mean of the average household incomes of the 21 towns?*

Answer to Sample Question 5.6.13 *$17,143*

Sample Question 5.6.14 *Is this the same as the average income of all the households in the 21 towns?*

Answer to Sample Question 5.6.14 *No way.*

Sample Question 5.6.15 *The custom test labelled* `MONEY` *is identical to what default test?*

Answer to Sample Question 5.6.15 *The t-test for* `INCOME`*.* $F = t^2$*, and the p-value is the same.*

Sample Question 5.6.16 *In the* `proc iml` *output, what can you learn from comparing* `F` *to* `FCRIT`*?*

Answer to Sample Question 5.6.16 $p < 0.05$

Sample Question 5.6.17 *For a town with 68,500 children 16 and under, and an average household income of $16,700, does the full model over-predict or under-predict sales? By how much?*

Answer to Sample Question 5.6.17 *Under-predict by $12,784. This is the first residual produced by* `proc print`*.*