Chapter 4

Multiple Linear Regression

Linear regression is the oldest and by far the most commonly used predictive modeling method. The biologist Sir Francis Galton, a cousin of Charles Darwin, is credited with its introduction in 1877 (Bulmer, 2003). He was actually less interested in prediction per se than he was in studying how traits were passed down from one generation to the next. He plotted characteristics of one generation against characteristics of the next, and noted that he could draw a straight line through the plot. Sweet peas that were heavier than average tended to produce offspring sweet peas that were heavier, tall fathers tended to have tall sons, and smart parents tended to have smart kids. However, what initially intrigued him was that the offspring of individuals with more extreme characteristics, that is, characteristics that are far from the mean of the population, tended to be closer to the mean than their parents. He referred to this as "regression to the mean," and since then we have been stuck with the rather uninformative name "regression" for the method.[1]

Since that time, a massive amount of statistical machinery has been developed to refine and extend Galton's seminal ideas. It is worthwhile to note that if we only have two variables—a single predictor variable and a target variable—this statistical machinery is overkill. We can do very nicely by using Galton's simple method of plotting the data. The reason for the algebra and associated number crunching is simply that there isn't much of interest that we can predict with only one predictor variable, and as soon as we get more than one, plotting goes quickly from impractical to impossible. With 3D plotting software we can visualize two predictors and one target, but that is the limit of conventional plotting in the three-dimensional world we live in. For anything else we need algebra, which is not limited to three dimensions.

The problem faced by those early statisticians was how to do with numbers what Galton did with his eyeball and his plots. Once that was figured out for one predictor and one target, it was relatively easy to generalize the approach to multiple predictor variables. The resulting method was called multiple linear

[1] "Regression to the mean" is a purely statistical effect and has nothing to do with genetics. In sports, an example of this is the so-called "sophomore slump." A player who has an outstanding rookie season must credit some part of the success to "luck" (really random chance). That luck can't be counted on in the next season, and so the player, now a sophomore, is likely not to do as well.

Table 4.1: Variable Roles

Statisticians say...	Data Miners say...	Meaning...
Independent Variables	Predictor Variables	Things you can use as inputs to predict an outcome
Dependent Variable	Target Variable	The outcome you want to predict with the inputs

regression, or more conveniently, *multiple regression*. While Galton understood the idea of multiple regression, he was unable to formalize the algebra. It was left to Karl Pearson, of Pearson Correlation Coefficient fame, and a colleague of Galton's, to master this tricky problem (Bulmer, 2003).

In our exposition, we will follow this historical logic, by first demonstrating the key concepts of regression using Galton's method of plotting one predictor against the target variable, and then translating those insights into the associated algebra. Most of the calculations necessary to generate numbers like coefficients and p-values will remain behind the scenes in our discussion, since today we can rely on numerically fluent computers to handle those details for us.

Once the foundations have been established, we will describe two useful extensions:

1. Incorporating categorical predictor variables
2. Handling nonlinear relationships with a linear model

We will also review the meaning and use of the coefficients, p-values, and R^2 that appear in the output of regression software.

4.1 Jargon Clarification

To start, we will clarify a bit of jargon, since statistical methods and data mining both developed in a number of different fields, with the different fields using different terms to describe the same concept. As an example, the terms describing the basic roles of variables differ across fields, as Table 4.1 illustrates.

We present the usual cautionary note that there is an implicit notion of causality here—the predictors are considered to cause the target outcome. However, causality is never proven by this method alone when applied to actual data. Only data from controlled experiments can allow us to conclude that a predictor causes the target. Regression on data that is not from controlled experiments can only indicate that changes in a target and a predictor variable occur together. Like chickens and eggs.

4.2 Graphical and Algebraic Representation of the Single Predictor Problem

Speaking of eggs, imagine a problem facing a supermarket category manager in southern California, who puts in orders for one week of eggs every Monday, and would like to predict the current week's volume of egg sales to know how much to order. The manager knows today's egg price per dozen, $1.05, and she has about two years of historical data available:

1. Target variable: number of cases of eggs sold each week
2. Predictor variable: weekly egg prices

Let's start by looking at the raw data. Some of these data are shown in Table 4.3. Even with only two variables, extracting useful information from a table of numbers is extremely difficult.

Let's try Galton's trick of plotting the numbers and see if things become more transparent. Figure 4.1 shows the scatterplot, and it does indeed look more informative. We can use this scatterplot to look up historical sales. Before reading on, try it—answer these two questions from the graph: If egg prices this Monday are at $1.05, what would you expect the weekly sales of eggs to be? What would you expect sales to be at a price of $0.80?

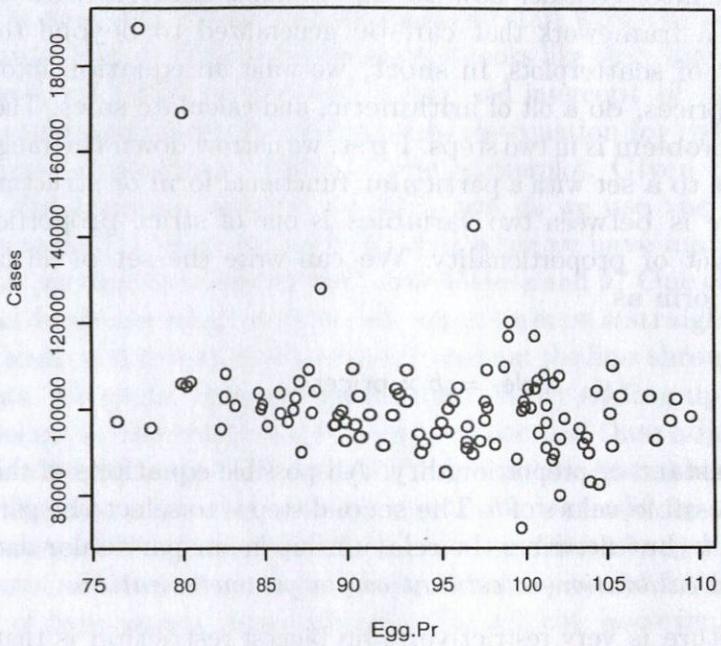
In coming up with sales figures, you have just used a predictive model. Congratulations! You probably came up with an estimate of around 103,000, plus or minus 5,000 for the first, and 95,000, plus or minus 15,000 for the second. Of course, you very likely also want to add some cautionary notes to those sales forecasts if it is to be used as a prediction for the upcoming week. Again, stop for a moment and make a mental list of some of those cautionary notes.

We will mention three things here.

Table 4.3: Egg Prices and Sales

Cases Sold/Week	Egg.Prices
96343	90.42
96345	89.33
96928	89.89
93519	90.71
99032	85.99
91539	91.83
89969	87.29
90859	96.36
99697	99.71
88350	99.38
100383	97.53
94415	100.77
91813	98.00
100466	99.89
96783	101.26
91008	101.26
100324	100.28
106628	100.69
98892	105.16
98252	109.55

Figure 4.1: Weekly Eggs Sales and Prices in Southern California



1. The relationship between price and sales for the coming week may not be the same as the past relationship. This problem always exists when using historical data as a basis for prediction, and the analyst needs to apply common sense and experience to judge if the past relations are likely to be relevant this week. Obviously, if the analyst is also a manager with experience in the industry, he or she will be in a good position to make that judgment.

2. The scatterplot indicates a range of possibilities for the prediction. Take a look at your sales prediction for \$0.80 and \$1.05, and this time include a range of possible values. A good way to think of the range is to think to yourself "what is the range of values that I would be 95% sure that the sales this week will be within?" Before reading on, look at the scatterplot and try this for \$0.80 and \$1.05.

3. Notice that within the scatter of the data points, egg prices don't seem to have too much effect on sales. The cluster of data points is rather flat. Granted, our estimate of sales is lower for higher prices, which is reasonable, but the range of sales at any price is large enough that we would not be wildly confident of lower sales with higher price in any specific case.

That was fairly easy, intuitive, and only took elementary school math skills. However, since our ultimate purpose is to tackle problems with many predictor variables, we will next consider how to do the same analysis with algebra, so that we have a framework that can be generalized to beyond the two-dimensional limit of scatterplots. In short, we want an equation into which we can plug egg prices, do a bit of arithmetic, and calculate sales. The usual approach to this problem is in two steps. First, we narrow down the range of all possible equations to a set with a particular functional form or structure. The simplest structure is between two variables is one of strict proportionality, with some constant of proportionality. We can write the set of all possible equations of this form as

$$\text{Sales} = b \times \text{prices},$$

where b is the constant of proportionality. All possible equations of this form are given by all possible values of b. The second step is to select the particular numeric value for b that describes the relationship in our particular data set.[2] This step is called *calibration*, or *estimation*, or *parameterization*.

This simple structure is very restrictive. The biggest restriction is that when prices are zero, sales must be zero. As such, it is not a very good model for this, or many other, situations. We can make it more flexible, and applicable to more situations, by adding another parameter, a, which will be the level of sales when egg prices are zero.

$$\text{Sales} = a + b \times \text{prices}$$

With all possible combinations of values of a and b, we now have a much larger set of possible relations between prices and sales. Again, we would use the data to *calibrate* (or estimate) the equation (i.e., to select specific values of a and b). Of course, once calibrated and we have specific values for a and b, we could plug in any value for price (such as \$1.05) and calculate the predicted value for sales. It turns out that if we plot the combinations of sales and price given by this equation for specific values of a and b, they fall on a straight line. Hence, we call this a linear model, and a linear model is the structural form assumed by linear regression. The new parameter a is called the *intercept*, because the line defined by the equation crosses, or intercepts, the sales axis at the value of a. The parameter b is the *slope* of the line. The linear model is of course

[2]By a "particular numeric value" we mean a number, like 2.45 or 763,457, rather than a symbol like b.

still restrictive, and many other structures in the relationships between two variables occur in the real world.[3]

If, as was the case with Galton's sweet peas, we start with the scatterplot and draw a straight line through the data, and it looks like that line approximates the data well, we could measure the slope and intercept of the line on the graph. We could substitute those values into the equation for the linear model, and now have an algebraic version of the scatterplot. Given prices, we can find sales. The question remains, though, how do we use the data, *without* making a scatterplot (since we can't do that when we have many predictors), to determine particular values for the parameters a and b? One obvious answer is to try and duplicate what we do by eye when we draw a straight line through a bunch of scattered points, which is to try and put the line through the middle of the points. We could duplicate this numerically by adding up the distances of all the points on the scatterplot from a trial line, and then adjusting the line so that this total distance is minimized. Other variations on this rule could be imagined. The particular rule in linear regression is to add up the the square of the distances of the points from the line in the y-direction, and to minimize that.[4] It turns out that surprisingly simple formulas can be found for a and b, given a set of data points, using this rule. We will not, however, go into those details. We will simply use R to take our data and calculate a and b using those formulas. The result is that a, the intercept or the expected sales level when the price is zero, is 153414 cases of eggs. The coefficient of prices, or the slope, is -554. The regression equation, which we can now use for prediction, is

$$\text{Sales} = 153414 - 554 \times \text{prices}$$

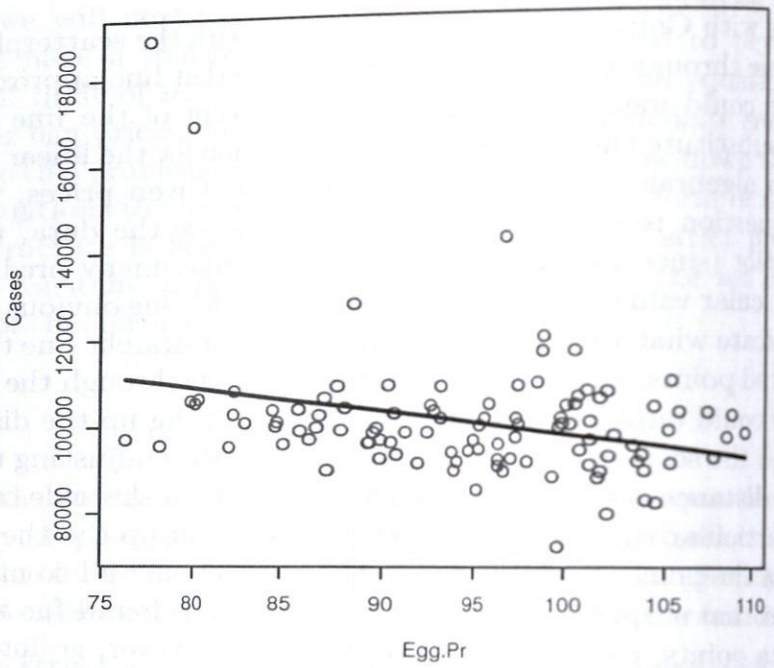
Note that the slope is negative, which means that as prices increase, sales decrease in this predictive model, as we suggested when we did our eyeball predictions. By the way, we should take that to mean that our numerical methodology seems to work because it duplicates our eyeballing, not that our eyeballing is confirmed by the method. Let's plot the values of sales and prices that satisfy this equation on top of our scatterplot, as in Figure 4.2.

Interestingly, this may not be the line you would have drawn by hand. If you agreed with the eyeball prediction estimates above, you likely would have

[3]Such as diminishing returns, increasing returns, U-shaped, and so on. We typically divide the universe into linear relationships and everything else, called nonlinear relationships. This is rather like dividing the zoological universe into pachydermic and non-pachydermic animals.

[4]To see an interactive animation that allows you to manually adjust the slope and intercept of a line to adjust the squared y-distance to a set of points, see http://www.dynamicgeometry.com/javasketchpad/gallery/pages/least_squares.php.

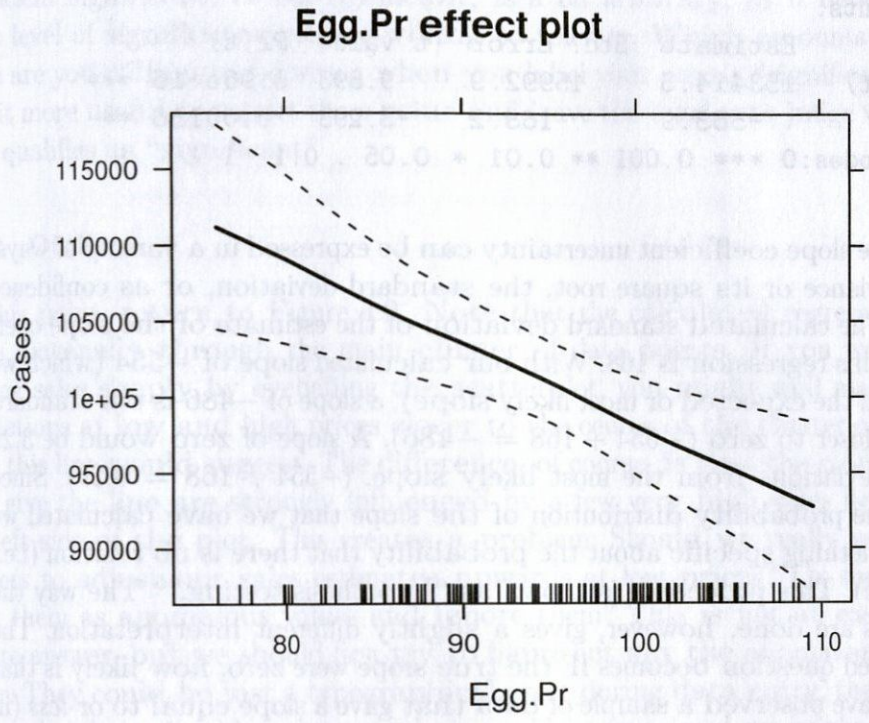
Figure 4.2: The Regression Line and Scatterplot



drawn a flatter line. Again, we should take that as a problem with the method, not with our eyeballing, since we are trying to find a numerical way to duplicate what we have easily done by eye graphically. We will come back to this in a minute. First, let's recall that when we used the scatterplot for sales predictions, we were also able to provide some sense of the possible range of the prediction. The formula, and the associated straight line, does not provide this useful information. Fortunately, the statistical machinery has also been developed for assessing this range, or variability. The machinery translates the scatter in the data to uncertainty in predictions of sales. Once again, we will not delve into the details of the calculations, but simply make the observation that if we can do this by eye, it should not be too surprising that we can find ways to do it with numbers. Just as we can show the calculated regression equation as a straight line to get a visual interpretation, we can plot the calculated uncertainty associated with the prediction line as error limits around that line (Figure 4.3) to provide a visual interpretation of the prediction error.

Although not quite the precise technical description of confidence limits, we can think of the prediction error shown here as a 95% chance that the actual realized case sales will fall between the dotted lines. This range is probably quite close to the range you guessed during the eyeballing exercise.

Figure 4.3: 95% Confidence Limits of the Regression Prediction



4.2.1 The Probability of a Relationship between the Variables

In many regression applications it seems that the main question of interest to the analyst is the rather weak one of whether the predictor variable is likely to be related to the target variable at all, given the data available. A slope coefficient of zero, corresponding to a horizontal regression line, means that there is no relation between the variables. That is, knowing the value of the predictor variable cannot refine your estimate of the target variable. Your best guess is always the mean of the target regardless of the value of the predictor. The scatterplot in Figure 4.1 suggests that we are close to the zero slope case for the relation between egg prices and sales. In Figure 4.3, however, the slope is quite steep. In fact, if we were to twist the line to horizontal some of it would have to be outside the 95% confidence limits for prediction, which suggests that it is highly unlikely that the true slope could be zero. This suggests asking a more refined question: "What is the probability that the true slope is actually zero?"

To answer this we need to know the range of uncertainty in the value of the slope coefficient. This uncertainty calculation again uses the scatter of the data. The more scattered the data, the greater the possible range of the coef-

Table 4.5: Regression Output for the Eggs Data

```
Coefficients:
             Estimate  Std. Error  t value  Pr(t)
(Intercept)  153414.5     15992.9    9.593  5.96e-16 ***
Egg.Pr         -553.9       168.2   -3.293   0.00136 **
Signif. codes:0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

ficients. The slope coefficient uncertainty can be expressed in a variety of ways, such as variance or its square root, the standard deviation, or as confidence intervals. The calculated standard deviation of the estimate of the slope coefficient for this regression is 168. With our calculated slope of −554 (which we can now call the expected or most likely slope), a slope of −486 is one standard deviation closer to zero (−554 + 168 = −486). A slope of zero would be 3.29 standard deviations from the most likely slope. (−554 / 168 = 3.29). Since we know the probability distribution of the slope that we have calculated we can say something specific about the probability that there is no relation (i.e., a zero slope). This procedure is known as "hypothesis testing."[5] The way the calculations are done, however, gives a slightly different interpretation. The actual refined question becomes IF the true slope were zero, how likely is that we would have observed a sample of data that gave a slope equal to or *less* (in this case) *than* the calculated mean slope of −544. From the *t*-distribution, this probability is 0.00136, or highly unlikely. While not technically correct, we can also think of this as indicating that it is highly unlikely that the true slope is zero, or highly like that there is indeed a relationship. All of the above information is contained in the regression printout in the row starting with "Egg.Pr," shown in Table 4.5. It is very worthwhile to work across this row and ensure that you understand the meanings of each of the numbers by relating then back to the preceding discussion and plots.

The stars beside the probability are a bit of a holdover from the days when distribution tables were used, and exact probability values were not calculated. They indicate the range in which the probability falls, and are interpreted in the last line as significance codes. Two stars, for example, means that the probability is between 0.01 and 0.001, and the categories are referred to as

[5]Or, more correctly, the people who programmed the software know the distribution. It was derived by William Gosset, a statistician and chemist employed by the Guinness brewery in Dublin from 1899 until his death in 1937. The story goes that Arthur Guinness was one of the earliest businessmen to recognize that skill in quantitative methods conferred a powerful competitive advantage, and did not want competitors to know that he was using them. He therefore prohibited employees from publishing their results. Hence, when Gossett published his results in 1908, he used the pen name "Student" (Pearson 1990). It was Fisher who labeled the statistic "Student's *t*." The rest is history.

statistical significance levels. From this, it should be obvious that calling a coefficient significant, or not significant, is a bit arbitrary, as it depends on what level of significance you are willing to tolerate. Which amounts to how often are you willing to be wrong when you label your result "significant." We find it more useful to report the p-value and leave the reader to judge whether that qualifies as "significant."

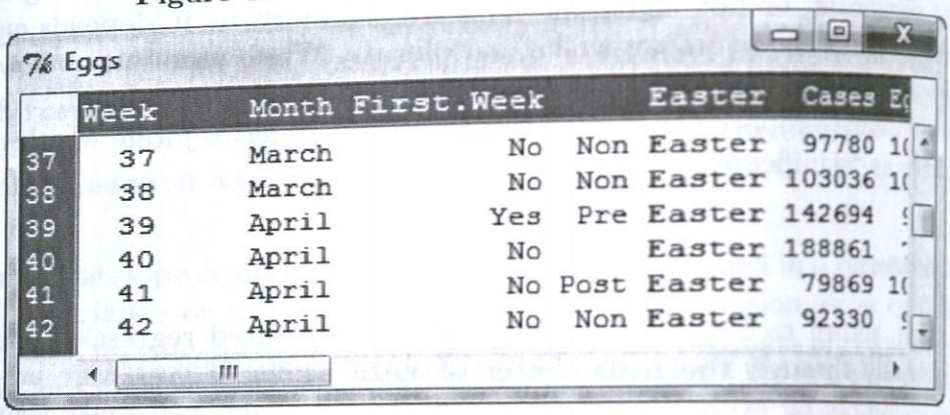
4.2.2 Outliers

At this point, return to Figure 4.2. Note that the calculated regression line slices diagonally through the main cluster of data points. If you were predicting sales simply by eyeballing the scatterplot, you might well make your predictions at low and high prices closer to the center of the cluster of points than this line would suggest. The difference, of course, is that the calculations that give the line are strongly influenced by a few very high sales figures on the left side of the plot. This creates a problem: Should we really use these outliers to adjust our sales estimates upwards at low prices? Or should we treat them as anomalous values and ignore them? This is not an easy question to answer, but we should first try to figure out why the anomalous values occur. They could be just a typographical error during data entry; they could be true but inexplicable values; or they could be true, and we might be able to figure out what caused them and use that information for future predictions. If we cannot come up with a reason for the outlier, it is usually best to drop it from the data set and re-estimate the regression, which will give us a prediction formula that is more like an "eyeballing" prediction that chose to ignore these large sales values. We will pursue the course of finding a reason for these outliers next.

4.3 Multiple Regression

For our eggs data, we also have other potential predictor variables available, and by exploring these variables a bit, we find that the outliers occur around Easter. Common sense says that means that we can make a prediction that starts with a price effect on sales which do not occur at Easter (in the eyeballing case, that means ignoring the outliers), and then, if our Monday eggs order is around Easter, make an adjustment to that prediction. Once we've identified when the outliers occurred, this is easy enough. To do the same numerically, we need to add at least one more predictor variable to the analysis, and we are now into the wonderful world of *multiple* regression. Graphical

Figure 4.4: The Data View for the Eggs Data Set



plots rapidly become overwhelmed, but the algebra we have developed can be extended quite easily.

4.3.1 Categorical Predictors

As long as our predictors are continuous (interval or ratio scale) tossing more of them into the regression algorithm and interpreting the outcome, is fairly straightforward.[6] "Easter," however, is a categorical variable. We can write

$$\text{Sales} = a + b \times \text{Price} + c \times \text{Easter}$$

and understand that we replace price with a number, like $1.05. But what do we replace Easter with? Furthermore, if we look at the actual data, we see that "Price" is a variable name at the top of a column that contains numeric values, but in the "Easter" column the values are "Non Easter, Pre Easter, Post Easter, and Easter" (Figure 4.4).

We cannot add, subtract, multiply, etc. "Non Easter," and so can't just throw the values into the number cruncher. In fact the "Easter" variable is nominal, and as we have seen in a previous chapter, you cannot do arithmetic on nominal variables. The way around this is to create new variables based on the Easter variable, which will indicate with a zero or one what week we are in. Our first indicator variable will indicate whether or not it is Easter, and we will call it "IndEaster," and assign it the following values:

$$\text{IndEaster} = 1 \text{ when Easter (the variable)} = \text{Easter (the value)},$$

[6] There is one major pitfall, namely correlated predictors, or *multicollinearity*, which we will not discuss in this chapter.

IndEaster = 0 for all other values of Easter

"IndEaster" contains numbers—numbers that our software can crunch—whose value depends on "Easter." The prediction formula now becomes

$$\text{Sales} = a + b \times \text{Price} + c \times \text{IndEaster}.$$

Note what happens once we have the coefficients estimated and go on to use this formula for prediction. When the week we are predicting is not Easter, the value of IndEaster is zero, and the last term (c x IndEaster) is zero. Sales will depend only on price. But if it is Easter week when we wish to predict sales, IndEaster is 1, and the last term takes the value we have estimated for the coefficient c. This means that the value of c is the sales boost that occurs during Easter. As a bonus, because the Easter outlier data points are now taken care of by IndEaster, they no longer affect the other coefficients (they don't pull the price line up at the low prices) so that the coefficient of price will be smaller, and the prediction regression line flatter.

We can add additional indicator variables for Pre Easter and Post Easter. Note that we have four values for the categorical variable, but we can only create three indicator variables. The coefficients are interpreted as the change in sales relative to the value for which we have not created an indicator variable, namely Non Easter. This is a general rule: If a categorical variable has n values, create $n - 1$ indicator variables, and interpret their coefficients as the change relative to the n^{th} variable.

When we estimate our new multiple regression equation, we get the following:

```
Coefficients:
                          Estimate        Pr(t)
(Intercept)             115387.19        2e-16 ***
Egg.Pr                    -170.15       0.0813
Easter[T.Pre Easter]     32728.55     1.94e-08 ***
Easter[T.Easter]         76946.67        2e-16 ***
Easter[T.Post Easter]   -22096.43     8.25e-05 ***
```

The coefficient estimates give the *sales prediction formula*

$$\text{Sales} = 115387 - 170 \times \text{Price} + 76946 \times \text{Ind Easter} + 32728 \times \text{Ind Pre Easter} - 22096 \times \text{Ind Post Easter}$$

We will note two things. First, if we are predicting egg sales during the Pre Easter week, we need to add 32,728 cases to the price-only prediction. If it is Easter, we add 77,000 cases, and if it is the week after Easter, everybody is sick of eggs, and we need to subtract 22,000 cases from our price-only prediction.

Second, now that we are explicitly considering the Easter season (some jargon: we say "controlling for Easter"), the slope of the price line is much shallower at −170. This is about one standard deviation away from zero. If the true value were zero, the chance of getting these data is 0.08 (about one in 12), which might cause some concern to the analyst who is asking, "Is there any relation at all between egg prices and egg sales?" This analyst might also say (more jargon) that the price coefficient "is significant at 0.1 level" since 0.08 is less than 0.1, but more than 0.05.

R Commander will also give you plots of the effect of one variable at a time, holding other variables constant, called *effect plots*. Figure 4.5 shows the effect plot for price with the new coefficients, as well as the 95% confidence intervals. Note that now the line could be horizontal or even positive and predictions from it would be within the confidence limits. Now that we have controlled for the Easter effect, we are much less certain that there really is a price effect on egg sales.

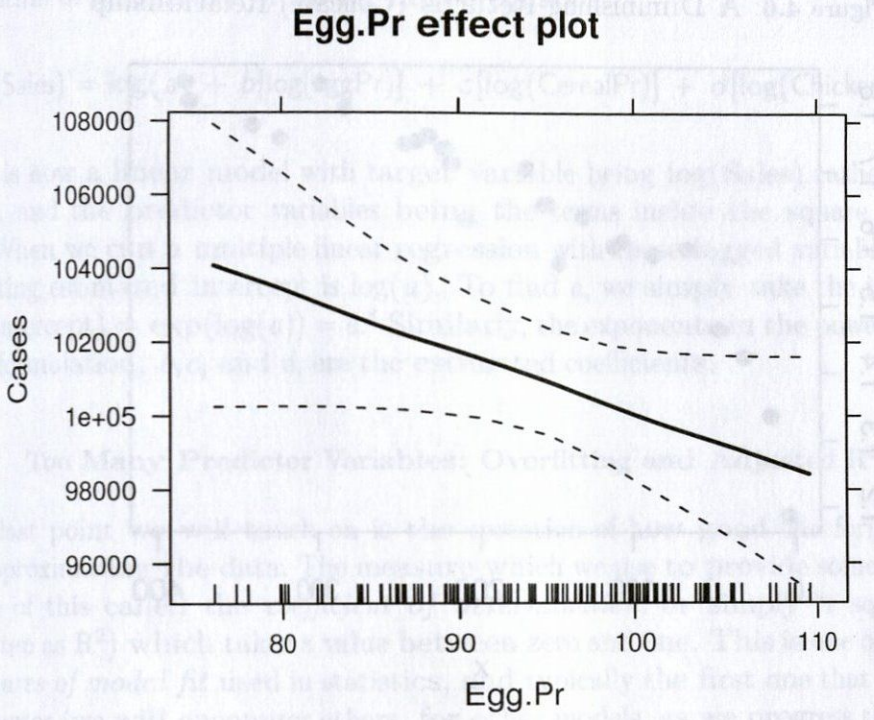
4.3.2 Nonlinear Relationships and Variable Transformations

At the beginning of this chapter, we said that the first step in translating our graph and eyeballing method to algebra was to decide on the model structure, or the functional form of the equation we will use. We then chose the linear model, and have spent all of our time since then dealing with the second step in the translation, namely *calibrating* the linear model. As noted in footnote 3, the world is awfully nonlinear. Nevertheless, the algebra of the one-predictor linear model seems really useful in extending the analysis to multiple predictors. We will give two responses to this little conundrum.

First, a linear model may be good enough. This is a judgment call based on how nonlinear the relationship is between each predictor variable and the target. We can use graphical visualization techniques to explore the relationships and decide whether any curves and twists we see in the data are big enough to justify the extra work of nonlinear modeling. In essence, we have to ask if approximating those curves and twists with a straight line will cause our prediction to be bad enough to cost us more from bad predictions than than the cost and effort of nonlinear modeling.

Second, we may have a relation between the variables that can be converted to a linear relationship by transforming the variables. A common nonlinear

Figure 4.5: Egg Price Effect Plot When Controlling for Easter



relationship between a predictor and a target is of the diminishing-returns variety, as shown in the stylized example in Figure 4.6.

If our graphical explorations of the data showed something like this, we would know that a straight line through those points would be a fair, but not great, approximation of the data. However, if we were to take the natural logarithm of the predictor variable (the one on the x-axis) and re-plot the results, the data points come closer to a straight line (Figure 4.7). Now, if we use a linear model on this transformed data, we can get a better fit, and better predictions.

Besides simply exploring relationships in the data graphically, we sometimes have theory or experience that tells us likely model structure. One common example of a nonlinear relationship that can be easily converted to linear is the power function model. For our eggs example, a power function model of egg sales as a function of egg prices, cereal prices, and chicken prices looks like

$$\mathsf{Sale} = a(\mathsf{EggPr})^b(\mathsf{CerealPr})^c(\mathsf{ChickenPr})^d,$$

where a, b, c, and d are the parameters we wish to estimate. We transform

Figure 4.6: A Diminishing Returns (Concave) Relationship

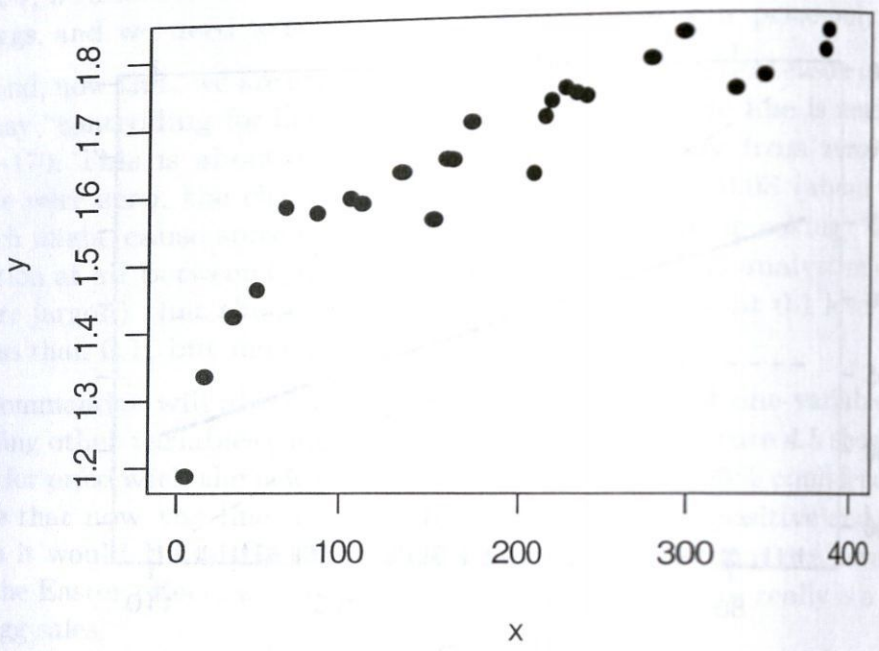
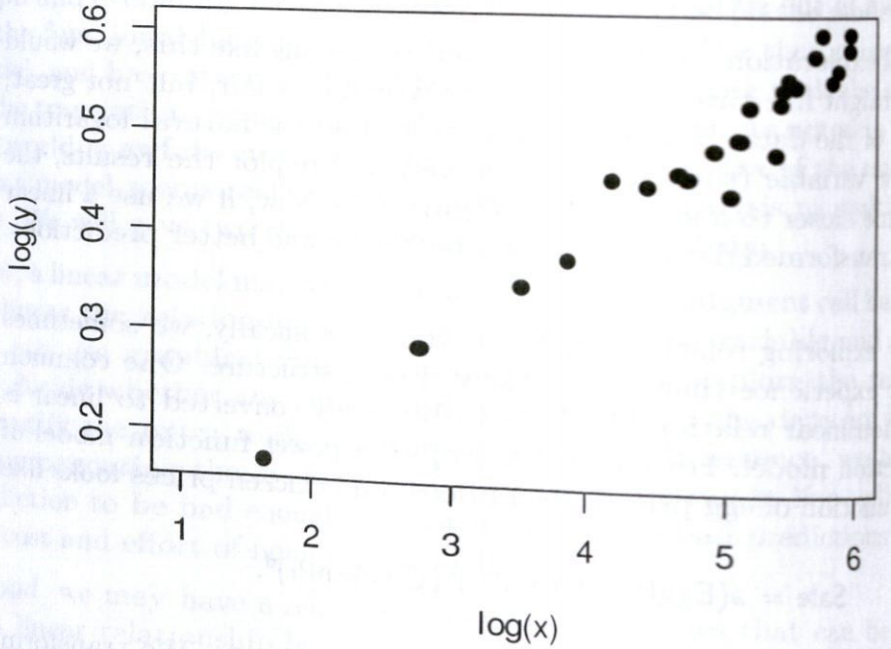


Figure 4.7: The Relationship after Logarithmic Transformation



this model into a form we can apply linear regression to by taking the natural logarithm of both sides,[7] resulting in the equation

$$\log(\text{Sales}) = \log(a) + b[\log(\text{eggPr})] + c[\log(\text{CerealPr})] + d[\log(\text{ChickenPr})].$$

This is now a linear model with target variable being log(Sales) rather than Sales, and the predictor variables being the terms inside the square brackets. When we run a multiple linear regression with these logged variables, the resulting estimated intercept is $\log(a)$. To find a, we simply take the inverse, $\exp(\text{intercept}) = \exp(\log(a)) = a$.[8] Similarly, the exponents in the power function formulation, b,c, and d, are the estimated coefficients.

4.3.3 Too Many Predictor Variables: Overfitting and Adjusted R^2

The last point we will touch on is the question of how good the formula is at approximating the data. The measure which we use to provide some initial sense of this called the *coefficient of determination*, or simply "r squared" (written as R^2) which takes a value between zero and one. This is one of many *measures of model fit* used in statistics, and typically the first one that people encounter (we will encounter others, for other models, as we progress through the book). Intuitively, the closer the data points lie to the regression line, the closer the R^2 value is to one.

When we have many possible variables to use, as is common in data mining, we would like to have some idea of which ones we should use and which ones we should not include. One might think that as long as the fit keeps improving as you add variables, you might as well keep adding variables. An interesting phenomenon, however, is that the more variables included in the regression, the higher the R^2 value is likely to be, *even if there is absolutely no relation between the additional variables and the target variable* (Judge et al., 1982, p. 601). The likely improvement in fit will occur because each new variable will be able to randomly account for some of the variation in the target variable. That means that R^2 by itself will not work as a means of deciding how many variables to include. One way to address this problem is to reduce the value of R^2 by a small amount every time an additional variable is added. Ideally, this amount should be roughly the amount we would expect the value of R^2 to

[7] R, like most statistical software, defines the function log() to be the natural logarithm with base e, a constant with a value of approximately 2.71828, while the function that takes the base 10 logarithm (the logarithm you most commonly worked with in high school math classes) is log10(). In the book, we adopt R's convention for referring to the natural logarithm function of x as $\log(x)$.

[8] $\exp(x)$ is notation for e raised to the power of x.

increase due to chance alone. The more variables included in the regression, the more we subtract from the calculated value of R^2. Statistical programs will all do this calculation, and the output is called *adjusted* R^2. Adding variables will never decrease the R^2, and will typically increase it simply by the small bit of fit improvement that a purely random relation will generate, but may decrease the *adjusted* R^2. If the increase in R^2 is very small, the adjusted R^2 will decrease with the addition of variables. In essence, the adjustment attempts to compensate for purely random fit improvements. An automatic way to decide on the number of variables, therefore, is to maximize adjusted R^2. This helps prevent fitting pure random noise, which we call *overfitting*. While maximizing adjusted R^2 is one approach, it may not be the best approach. There is both a technical and practical reason for this. From a technical perspective it has been shown that adjusted R^2 does not actually adjust enough as additional variables are added to a model, so maximizing adjusted R^2 is not a guarantee against overfitting (Amemiya, 1985; pp. 49–51), but it is a lot better than relying on R^2 alone. The practical reason is that we also need to be concerned about whether the variables included in a model make sense from a managerial perspective, rather than including them purely because of their impact on adjusted R^2. We will have more, much more, to say about overfitting in later chapters.

4.4 Summary

1. If you have learned the terms Independent and Dependent, start getting used to **Predictor and Target.**
2. **Graphical and Algebraic Representation**, or who needs algebra? You do, but only if there is more than one thing affecting the target you are trying to predict.
3. The p-values give you an indication of how probable it is that there is no relationship between the predictor and target variable. This is called **statistical significance. It does not tell you anything about the size or importance of the effect.**
4. Problems and Solutions:
   - There is a lot of stuff printed out in the regression output.
     - The most important pieces of information for now are the coefficient values, their p-values, and adjusted R^2.

- Help! I've got a categorical predictor variable!
  - RELAX. We'll use **indicator variables**, aka dummy variables.
- The real world is awfully **nonlinear.**
  - Transform your variables, on the basis of theory, your experience, or explorations of the data.

4.5 Data Visualization and Linear Regression Tutorial

This tutorial has two purposes. First, it demonstrates some of the data visualization methods provided by R for exploring the relationships between a *continuous* (interval or ratio scaled) *target* (dependent) variable and *continuous* or *categorical predictor* (independent) variables. In addition, these plots can suggest the shape of the relationship, such as linear or diminishing returns. The second purpose of this tutorial is to provide an opportunity for you to become familiar with estimating linear regression models, including handling categorical predictors, and one common type of nonlinear model structure.

1.

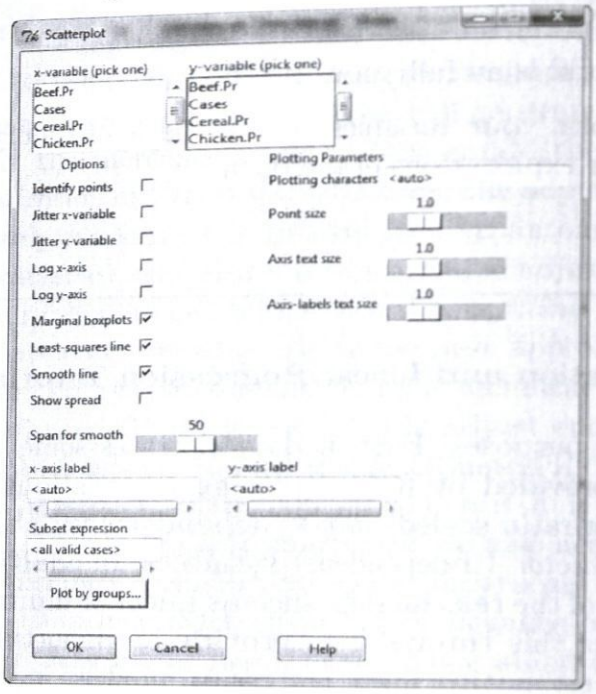
For this tutorial we will be using the Eggs data set from the BCA data library, which can be accessed with **Data → Get From → R Package → Read data set from an attached package...** pull-down menu option (described in more detail in previous tutorials). The Eggs data set provides information on weekly retail sales of eggs in southern California over a two-year period. Our objective is to determine what factors influence egg sales so that we (in the role of a retailer) can make better decisions regarding our pricing, ordering, and inventory decisions for eggs. You can view the R help file describing this data set using the pull-down menu option **Data → Clean → Help on active data set (if available)**....

2.

Scatterplots

We are going to begin by graphically examining the relationship between egg prices and egg sales using a scatterplot. To do this use the pull-down menu option **Explore and Test → Visualize →Scatterplot...**, which will bring up the dialog box shown in Figure 4.8. Since a scatterplot is used to examine continuous (ratio and interval) variables, those variables in the data are

Figure 4.8: Scatterplot Dialog



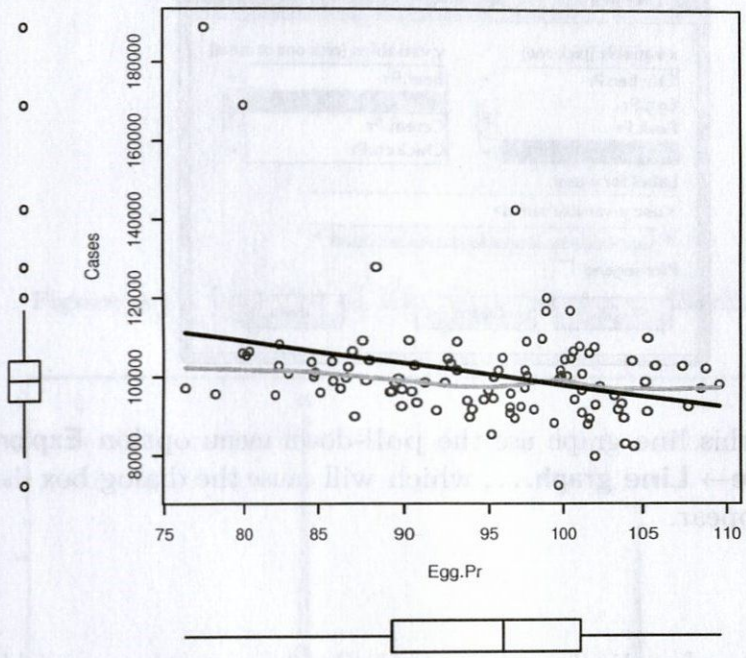
available in the dialog. **Select Egg.Pr (egg prices), the predictor, as the x-variable, and Cases (weekly egg case sales), the target variable, as the y-variable**. In addition, **check "Marginal boxplots," "Least-squares line," and "Smooth Line" boxes**. The "Marginal boxplots" option places a box plot of each of the two variables along each variable's axis, which gives a sense of the distribution of each of the variables in the plot. The "Least-squares line" and "Smooth Line" produce a plot of the fitted relationship between the two variables. The "Least-squares line" is the calibrated simple ("simple" means only one predictor, in contrast to "multiple") linear regression model, where `Cases` is the target (or "dependent") variable and `Egg.Pr` is the predictor (or "independent") variable. "Smooth Line" captures potential non-linear relationships in the data (the specifics of the method used to estimate the "Smooth Line" are beyond the scope of this book).[9]

3.

Once you have made the needed selections, **press OK**. A graphics window with the scatterplot will appear (Figure 4.9). An examination of the scatterplot points indicates that there appears to be a slight negative relationship

[9]For the curious, it is created using a "local regression" model known as *loess* (Cleveland and Devlin, 1988).

Figure 4.9: The Scatterplot of Eggs Sales and Prices



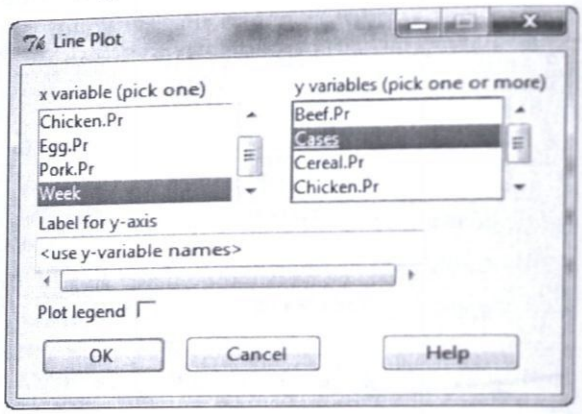
between egg sales and egg prices, and more noticeable negative slope on the least-squares (dotted) line. It also appears that the outliers are driving this negative slope. The negative relationship is less clear based on the smoothed (solid) line. The other use of the smoothed line is to look for evidence of non-linear relations between the two variables. If we see any simple, systematic curvature across the whole plot, we would likely try some nonlinear transformations of the variables, and then rerun the regression to improve the regression fit and capture the nonlinearity. In this case, there is very little evidence of any systematic curvatures, so we would be unlikely to attempt any transformations. Very shortly we will discuss how to interpret the box plots along the axes. First, let's see if we can find a reason for the outliers. If we cannot explain them we may want to drop them out of the analysis.

4.

Line Graphs

Another way to explore and visualize the relation between two variables is to create what is called a "line graph" in R Commander, between `Cases` and `Week`. Line graphs make sense *if the data is sorted* by one of the variables used in the analysis, which turns out to be true for the variable `Week` in this instance. This is often particularly useful where the sorted variable represents

Figure 4.10: Line Plot Dialog



time. To create this line graph use the pull-down menu option **Explore** and **Test** → **Visualize**→ **Line graph. . .** , which will cause the dialog box shown in Figure 4.10 to appear.

5.

In this dialog box select `Week` as the x variable, `Cases` as the y variable, and press **OK** to produce the plot shown in Figure 4.11. We can immediately see the same pattern of outliers appear about 52 weeks, or one year apart, and therefore represent something seasonal happening with egg sales. This then suggests that we should look back at the data to see if we can tell what the seasonal effect is.

6.

View the data (**click on the View Data Button**) to inspect what else is happening during the outlier weeks 40 and 91. Happily for us, the Eggs data set includes the variable `Easter`, a nominally scaled categorical variable (or "factor") that captures the "Easter effect" by identifying Easter week, the week before Easter, and the week after Easter. The two very pronounced spikes in weeks 40 and 91 shown in Figure 4.11 are a result of the Easter holiday, where there is a tradition of coloring hard boiled eggs for children on Easter. Comparing the figure with these values in the data reveals that sales of eggs start a steep climb the week prior to Easter Sunday, peak the week containing Easter Sunday, and then through the week following Easter Sunday. Our visual analysis has made an interesting discovery. This is, of course, a simple example, and an egg manager would already know this and not need to go through this analysis. For us, starting with simple examples

Figure 4.11: Line Plot of Egg Case Sales over Weeks

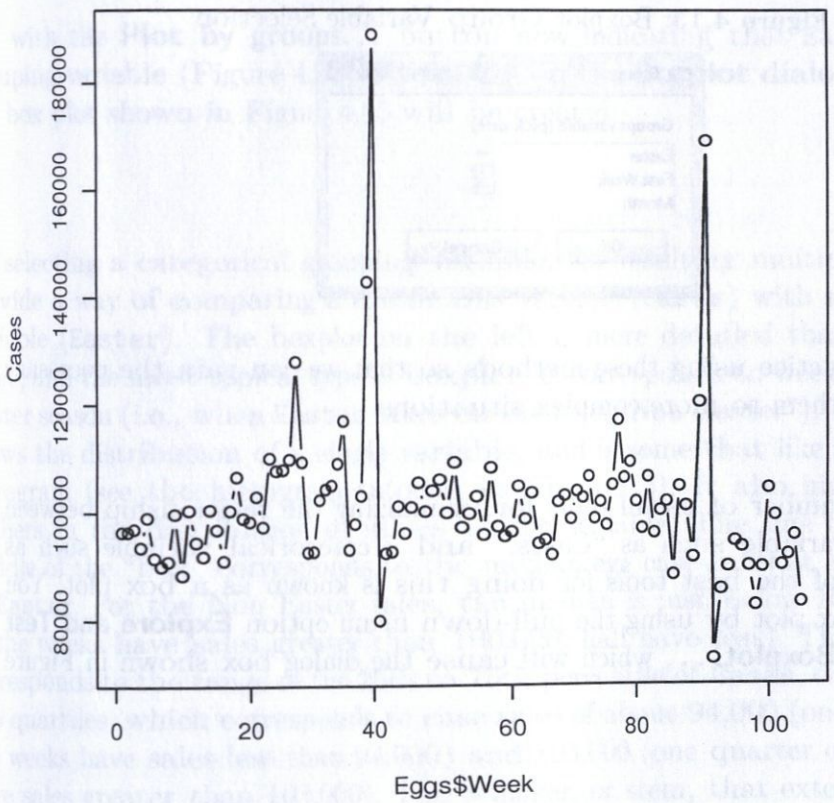


Figure 4.12: Boxplot Dialog

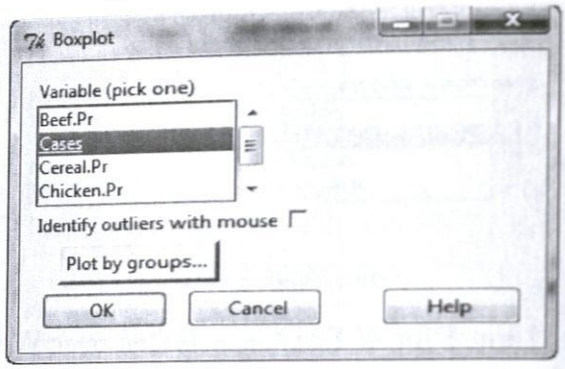
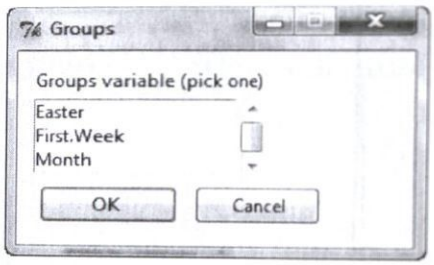


Figure 4.13: Boxplot Group Variable Selection



allows us to practice using these methods so that we can gain the necessary skills to apply them to more complex situations.

Boxplots

R provides a number of useful tools for visualizing the relationship between a continuous variable such as "Cases," and a categorical variable such as "Easter." One of the best tools for doing this is known as a box plot. You can create a box plot by using the pull-down menu option **Explore and Test** → **Visualize**→**Boxplot...**, which will cause the dialog box shown in Figure 4.12 to appear.

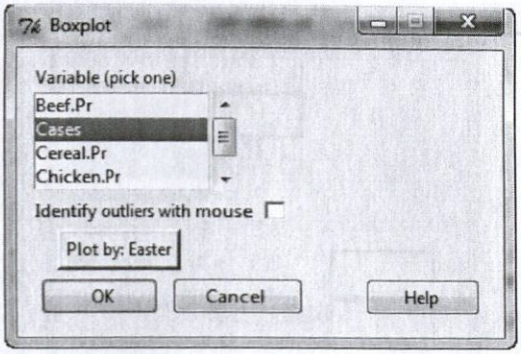
7.

In the **Boxplot** dialog box **select Cases as the "Variable (pick one)**," and then click on the **Plot by groups...** button, which will cause the dialog box shown in Figure 4.13 to appear, with any categorical variables that are in the data set that could be used as grouping variables.

8.

Select Easter as the "Groups variable (pick one)" in the **Group** dialog box, and then press **OK**. This will bring you back to the **Boxplot** dialog box,

Figure 4.14: The Revised Boxplot Dialog

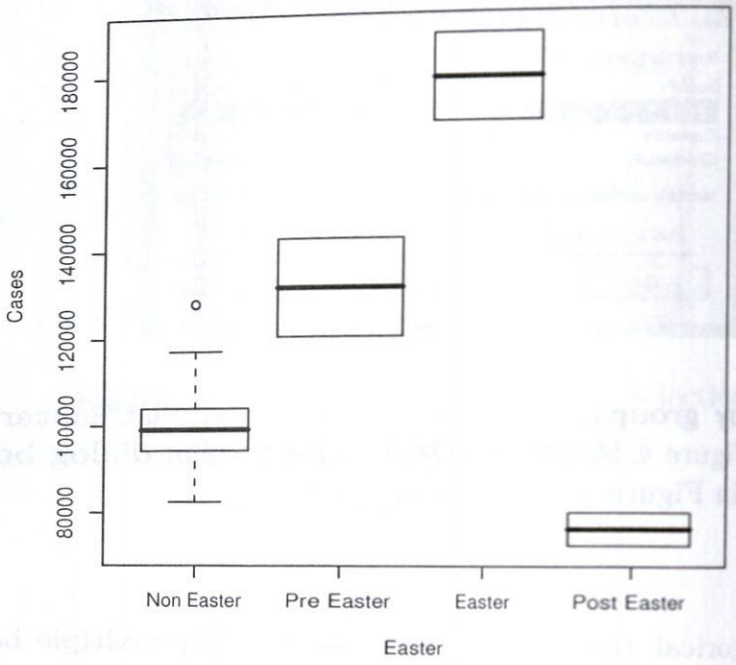


but with the **Plot by groups...** button now indicating that `Easter` is the grouping variable (Figure 4.14). **Press OK in the Boxplot dialog box**, and the box plot shown in Figure 4.15 will be created.

9.

By selecting a categorical grouping variable, the resulting multiple boxplots provide a way of comparing a *continuous* variable (`Cases`) with a *categorical* variable (`Easter`). The boxplot on the left is more detailed than the other three, and the most typical type of boxplot. It corresponds to weeks not in the Easter season (i.e., when `Easter` takes on the level "Non Easter"). The boxplot shows the distribution of a single variable, and is somewhat like a simplified histogram (see the histogram tutorial in Chapter 3). It also highlights the outliers in the distribution of values of the variable. The line towards the middle of the "box" corresponds to the median egg case sales at a given level of `Easter`. For the Non Easter sales, the median is just below 100,000 (half of the weeks have sales greater than 100,000, half have less). The main box corresponds to the range of the 25th to 75th percentile of `Cases`, or the middle two quartiles, which corresponds to case sales of about 94,000 (one quarter of the weeks have sales less than 94,000) and 103,000 (one quarter of the weeks have sales greater than 103,000). The whisker, or stem, that extends beyond the box goes out to the largest and smallest data values, unless those values are more than 1.5 times the width of the box. The height of the box corresponds to what is known as the "interquartile range" (greater than the lower quarter, less than the upper quarter). For Non Easter, the lowest sales are at about 82,000 cases. Any points that are more than 1.5 times the interquartile range away are represented by individual circles. The highest sales for Non Easter weeks is about 128,000 cases, which is represented by the small circle in the boxplot. This is the only week in which sales are more than 1.5 times the box width (interquartile range). The box plots for the Pre Easter, Easter, and Post

Figure 4.15: Boxplot of Egg Case Sales Grouped by Easter Weeks



Easter values of the categorical `Easter` variable are unusual in this instance in that they don't have whiskers. The reason for this is that there are only two data points for each. With only two years of data, we only observe two Easter seasons, and hence a 25th to 75th percentile range is meaningless. The median lines, however, are meaningful, and clearly demonstrate the enormous effect that the Easter season has on egg sales. Sales during these three weeks lie far outside the usual range of sales for non-Easter weeks.

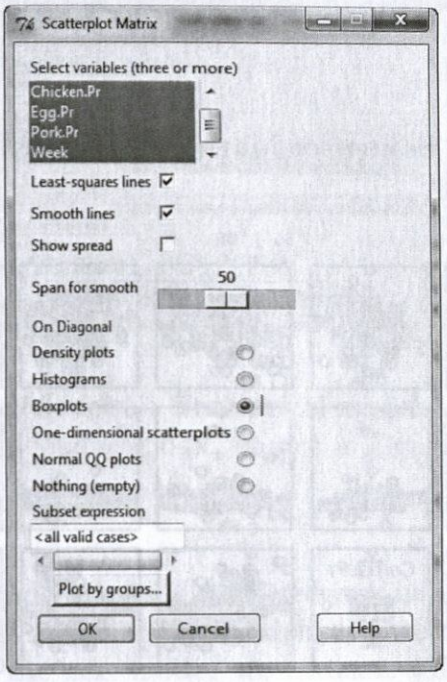
10.

Exercise: At this point you should check the relationship between the other variables included in the data set and case sales of eggs. Starting with the categorical variables, create a box plot of `Cases` for different levels of the variable `First.Week` (a variable that indicates whether an observation corresponds to the first week in a month). What are the median and quartile values (approximately) shown in the boxplots? Does `First.Week` appear to have an effect? What is that effect? Can you think of a plausible reason why this effect might exist?

Scatterplot Matrix

A tool that allows a quick exploration of the relationship among several continuous variables is the Scatterplot Matrix. **Select Explore and Test → Vi-**

Figure 4.16: Scatterplot Matrix Dialog



sualize → **Scatterplot Matrix...** to bring up the dialog box in Figure 4.16. **Select all seven of the continuous variables, and the boxplot option. Click OK.** The result in the graphics window will be a matrix of scatterplots for all combinations of the seven variables (Figure 4.17). This is close to the maximum number of variables that can be easily interpreted in the scatterplot matrix. The diagonal shows the boxplots for each variable. Horizontally along a row from the boxplot shows the same variable as the y-variable in the scatterplots; for example, the second row has Cases as the y-variable, with the other six variables as x-variables. The second row and fifth column, for example, shows the Egg prices vs. Cases scatterplot we examined earlier.

Exercise: Identify any other variables related to Egg Case sales, and speculate as to why the relation(s) might exist.

11.

Linear Regression

Having looked at the data visualization tools for continuous dependent variables, we now move on to an examination of R's linear regression tools. To estimate a multiple (more than one predictor) linear regression model use the

Figure 4.17: Scatterplot Matrix of the Eggs Data

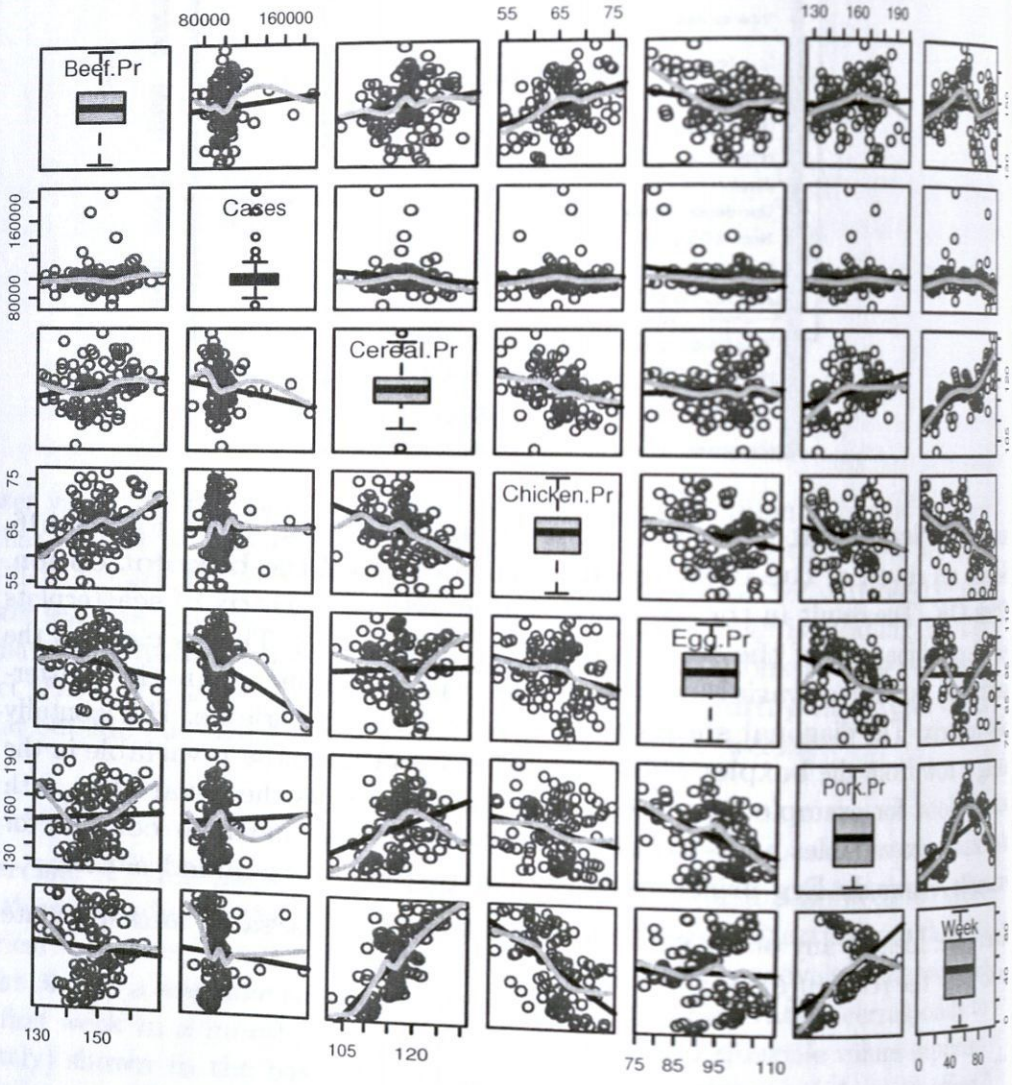
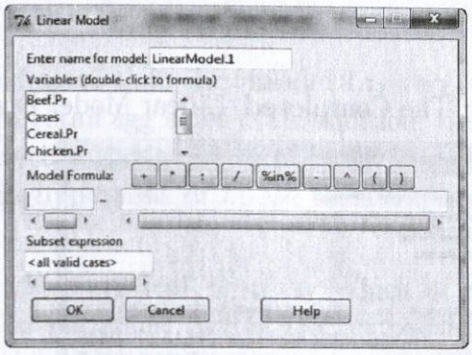


Figure 4.18: The Linear Model Dialog



pull-down menu option **Models** → **Statistical models** → **Linear model...** to bring up the Linear Model dialog box, shown in Figure 4.18.

12.

By default, R Commander names a linear regression model `LinearModel.`n, where n is the sequential number of models estimated during that R Commander session (thus the next model we estimate will be given the default name `LinearModel.2`). This is a very useful feature of R Commander, since it prevents us from writing over existing models by mistake. However, we are likely to prefer a more descriptive name. Since in the first model we will not be transforming any of the variables, the model we estimate will be truly "linear." Consequently, it makes sense to name this model LinearEggs since it indicates both the structure and the data set used in constructing the model, and **enter it in the "Enter name of model:"** field. Next **double-click on the variable `Cases`** in the "Variables (double-click to formula)" scroll box. By default, the first variable you double-click from this scroll box is selected as the dependent variable of the model. **Next double-click, in turn, the variables `Beef.Pr`, `Cereal.Pr`, `Chicken.Pr`, `Easter`, `Egg.Pr`, `First.Week`, `Month`, `Pork.Pr`. Do not select `Week`.** You have to select each variable individually, and cannot select multiple variables at once. Each variable will automatically be separated by a plus sign (+). R Commander also recognizes categorical (factor) variables like Easter and month, and automatically creates the necessary indicator (dummy) variables. When you have completed all of this, your dialog box should look like the one shown in Figure 4.19. When it does, **press OK**, and the results of the estimated model will appear in R Commander's output window (Figure 4.20). This window shows the usual regression ouput information, including coefficient estimates, t- and p-values, and the goodness of fit statistics R^2 and adjusted R^2.

Figure 4.19: The Completed Linear Model Dialog

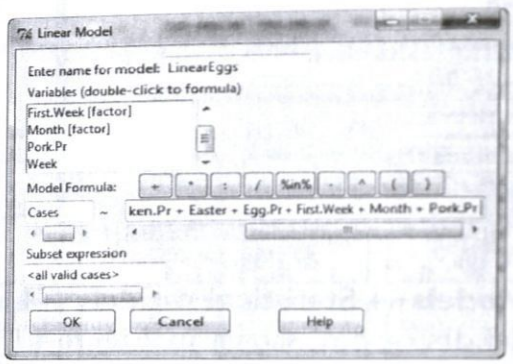


Figure 4.20: Linear Regression Results for LinearEggs

Output Window

Submit

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     165855.71   26796.27   6.190 2.12e-08 ***
Beef.Pr            283.38      96.36   2.941 0.004229 **
Cereal.Pr         -387.13     169.61  -2.282 0.024989 *
Chicken.Pr        -149.18     159.94  -0.933 0.353622
Easter[T.Pre Easter]  29548.19  4630.88   6.381 9.19e-09 ***
Easter[T.Easter]  73067.62    5274.35  13.853  < 2e-16 ***
Easter[T.Post Easter] -17113.09  4970.42  -3.443 0.000899 ***
Egg.Pr             -452.62     113.72  -3.980 0.000146 ***
First.Week[T.Yes]  5590.40    1396.16   4.004 0.000134 ***
Month[T.February] -2549.12    2816.69  -0.905 0.368050
Month[T.March]    -3614.77    2946.54  -1.227 0.223331
Month[T.April]   -10495.24    3413.45  -3.075 0.002843 **
Month[T.May]     -12730.61    2908.14  -4.378 3.43e-05 ***
Month[T.June]     -9811.37    2911.22  -3.370 0.001137 **
Month[T.July]    -11647.59    2631.78  -4.426 2.87e-05 ***
Month[T.August]  -15544.53    2793.86  -5.564 3.08e-07 ***
Month[T.September] -8912.41   2876.59  -3.098 0.002648 **
Month[T.October]  -9723.55    2835.23  -3.430 0.000939 ***
Month[T.November] -3648.33    2776.16  -1.314 0.192368
Month[T.December]  1327.28    2867.30   0.463 0.644630
Pork.Pr             -27.75      46.34  -0.599 0.550839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5691 on 84 degrees of freedom
Multiple R-squared: 0.8725, Adjusted R-squared: 0.8421
F-statistic: 28.74 on 20 and 84 DF,  p-value: < 2.2e-16
```

13.

Interpretation

In Figure 4.20, R^2 indicates that the model fits very well (it explains over 87% of the variance in retail egg sales). In addition, the results also indicate that many of the included variables are highly statistically significant based on their *p*-values (the null hypothesis of these tests is that the coefficients equal zero).

Exercise: First use the statistical results: Which of the variables seem to be most important? In particular, what do the coefficients on the monthly dummy variables suggest about the seasonal pattern of egg sales? What is the impact of an increase in the price of eggs on retail egg sales? An increase in the price of beef? An increase in the price of breakfast cereal? Now use your judgment: Do all of these results make logical sense? That is, are the signs (+ or −) and magnitudes of the coefficients consistent with your expectations?

14.

Interpreting Categorical Predictors

One issue with variables that are factors (categorical) is that each *level* of a factor has its own estimated coefficient that measures the impact of that level relative to the (omitted) "base case" level of that factor. For instance, for `Month` the omitted level is "January," and the parameters for the other 11 months indicate the *difference* (and whether it is statistically significant) between that month and January. For example, 15,544 fewer cases, on average, are sold in August than in January. As a result, different choices of the omitted level will alter the pattern of estimated coefficients and significance levels, although the choice of the omitted level has no influence on overall model fit or, if we use the model for prediction, the resulting predicted values of `Cases` sold. If we look across the levels of `Month` in the estimated model we find that some of the coefficients are negative, while others are positive. Similarly, some of the coefficients (and hence the difference in the effect from January) are statistically different from zero, while others are not. Finally, based on these results, we do not know, taken as a whole, whether `Month` is statistically significant. R Commander provides a very nice tool to assess whether, taken as a whole, a factor variable is statistically significant. To use this tool use the pull-down menu option **Assess →Hypothesis tests →ANOVA table**, and press **OK**, which will cause the results shown in Figure 4.21 to be printed to the R Commander output window. Examining the results in Figure 4.21 you can see that, taken as a whole, all three factor variables (`Easter`, `First.Week`, and `Month`) in this model are all highly statistically significant.

Figure 4.21: ANOVA Table Hypothesis Test

Output Window

```
> Anova(LinearEggs, type="II")
Anova Table (Type II tests)

Response: Cases
              Sum Sq Df F value    Pr(>F)
Beef.Pr    280103188  1  8.6478 0.0042295 **
Cereal.Pr  168742668  1  5.2097 0.0249886 *
Chicken.Pr  28180179  1  0.8700 0.3536222
Easter    8773987097  3 90.2948 < 2.2e-16 ***
Egg.Pr     513078180  1 15.8406 0.0001457 ***
First.Week 519311341  1 16.0330 0.0001338 ***
Month     2107016293 11  5.9137 4.886e-07 ***
Pork.Pr     11618467  1  0.3587 0.5508388
Residuals 2720772383 84
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

15.

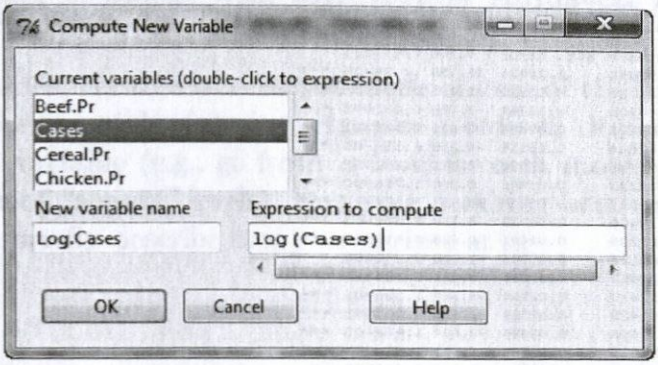
Nonlinear Model

Having estimated the linear model specification, we will next estimate the multiplicative power function specification. This specification assumes that, instead of `Cases` being a function of the sum of the predictor variables multiplied by a coefficient (that is estimated), `Cases` is the product of the predictor variables, each raised to the power of an estimated coefficient. This is easy to do, because if we take the natural logarithm of each side of the multiplicative power function model, it becomes a linear model of logs of the variables, and we can use standard linear regression to estimate the coefficients. To estimate the power function specification we must, therefore, apply a natural logarithm transformation to the dependent variable (`Cases`) and all the continuous (ratio or interval scaled) predictor variables (i.e., the five price variables). However, we do not apply the natural logarithm transformation to the three factor variables. In order to create new variables based on the appropriate variable transformations use the pull-down menu option **Data** → **Manipulate variables** → **Compute a new variable...**, which will bring up the dialog box shown in Figure 4.22.

16.

In the "New variable name" field enter the name you wish to give to the log transformed `Cases` variable, which can be any name you like. In Figure 4.22 the name `Log.Cases` has been typed in. The next box specifies the computation, which is to take the logarithm of each value of the Cases variable, and has to be

Figure 4.22: Compute New Variable Dialog



entered exactly as shown. Enter the formula `log(Cases)` into the "Expression to compute" field, and then press **OK** to create the new variable `Log.Cases`.

17.

Repeat steps 15 and 16 using the five price variables (changing variable names) to create the five natural log transformed price variables. Check that all of the variables have been created by **View Data**. One you have done this, repeat the regression analysis, only this time using the new log transformed variables rather than the original case and price variables (reviewing step 12 may help here). Specifically, this model should use the log transformed `Cases` and price variables along with the original three factor variables. This is equivalent to estimating the power function model of the original variables. One thing to be aware of is that when you start, the **Linear Model** dialog box will have your previous model initially entered. This is often useful since a common thing to do during an analysis is to add or remove variables from the previous model estimated. Although, in this case we are altering so many variables that it makes more sense to delete everything in the two model fields and start over. Once you have estimated the model, your results should look like those contained in Figure 4.23.

18.

Compare the results of the two estimated models. **Which fits the data better?**[10] **Exercise:** How do the estimated coefficients differ across these two models, particularly their signs and significance? Although the magnitudes are

[10] A technical point is that the adjusted R^2 values for the logged and unlogged dependent variables are not strictly comparable; however, the differences are usually minor. In subsequent chapters we introduce a different and better method of comparing models using holdout samples, which avoids the problem.

Figure 4.23: Linear Regession Results for the Power Function Model

Output Window

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             13.90880    1.18826  11.705  < 2e-16 ***
Log.Beef.Pr              0.39135    0.13520   2.895 0.004836 **
Log.Cereal.Pr           -0.40900    0.18764  -2.180 0.032076 *
Log.Chicken.Pr          -0.06125    0.09886  -0.620 0.537223
Log.Egg.Pr              -0.42666    0.10114  -4.218 6.18e-05 ***
Log.Pork.Pr             -0.03139    0.06936  -0.452 0.652098
Easter[T.Pre Easter]     0.26223    0.04401   5.958 5.77e-08 ***
Easter[T.Easter]         0.52815    0.05060  10.438  < 2e-16 ***
Easter[T.Post Easter]   -0.19409    0.04712  -4.119 8.86e-05 ***
Month[T.February]       -0.02205    0.02681  -0.823 0.413119
Month[T.March]          -0.03116    0.02806  -1.110 0.270009
Month[T.April]          -0.11087    0.03241  -3.421 0.000966 ***
Month[T.May]            -0.12647    0.02756  -4.588 1.55e-05 ***
Month[T.June]           -0.09423    0.02759  -3.416 0.000983 ***
Month[T.July]           -0.11508    0.02499  -4.604 1.46e-05 ***
Month[T.August]         -0.15438    0.02661  -5.802 1.13e-07 ***
Month[T.September]      -0.08326    0.02734  -3.046 0.003102 **
Month[T.October]        -0.08994    0.02704  -3.326 0.001308 **
Month[T.November]       -0.03067    0.02636  -1.164 0.247888
Month[T.December]        0.01695    0.02723   0.623 0.535248
First.Week[T.Yes]        0.05584    0.01327   4.207 6.46e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05409 on 84 degrees of freedom
Multiple R-squared: 0.8396, Adjusted R-squared: 0.8014
F-statistic: 21.99 on 20 and 84 DF,  p-value: < 2.2e-16
```

very different between the two models, are the larger magnitudes in one model the same as the larger magnitudes in the other? Which variables appear to have effects that are "robust" to the selection of a particular functional form (that is, remain fairly similar in sign and relative magnitude and significance in both models)?

19.

A Good Model

Exercise: Using these results, do some exploratory data analysis, with the **objective of estimating a model which minimizes the number of predictor variables used in the model, but retains a fairly high R^2.** For continuous predictors, try omitting non-significant predictors. The R^2 will typically drop with each variable removed, but will drop more with some variables than with others. An additional statistic reported that helps with the decision on which variables to retain and which to drop is Adjusted R^2. This number is calculated by subtracting an amount from R^2 which depends on the number of variables used in the regression. The more variables used, the more the R^2 is reduced by the adjustment. Adding variables will almost always increase the R^2, simply by the small bit of fit improvement that a purely random relation will generate; but may have little effect on the Adjusted R^2. If the increase in R^2 is very small, the Adjusted R^2 will decrease with the addition of variables. In essence, the adjustment attempts to compensate for

purely random fit improvements, but, unfortunately, does not completely do so. **Still, finding a model with the largest Adjusted R^2 is a good approach.**

For factor variables, try to relabel factor levels of some of them using the tools introduced in the tutorials of the last chapter in order to decrease the number of levels of the variable (e.g., go from a level for each month in a year to a smaller number of "season" levels). Put your resulting best model in a word document, and briefly describe how you arrived at it.