Regression with Big Data

THE NEWS IS FULL OF REFERENCES TO "BIG DATA." What does that mean, and what does it have to do with statistics? In business, big data usually refers to information that is captured by computer systems that monitor various transactions. For instance, cellular telephone companies carefully watch which customers stay with their network and which leave when their contracts expire. Systems that monitor customers are essential for billing, but this information can also help managers understand customer churn and devise plans to keep profitable customers. Traditional bricks-and-mortar retail stores also generate wide streams of data. Modern checkout software tracks every item a shopper buys. These systems were originally devised to monitor inventory, but are now used to discover which items shoppers buy. That information can help marketers design coupons that entice shoppers to buy more on their next visit.[1]

Big data can be overwhelming. Imagine opening a data table with millions of rows and thousands of columns. That's a lot of numbers, and not all of that information is typically useful. Where do you begin? Transactional data often aren't relevant for building a statistical model. Figuring out what to use and what to ignore is a challenge, and we can only expect to succeed by first understanding the business. Unless we have a clear sense of the problem and the context, we're not going to be able to sort through thousands of columns of numbers to find those that help us. Automated modeling tools offer help (see Statistics in Action: Automated Modeling), but we can do better if we understand the problem.

THIS CHAPTER PRESENTS A SYSTEMATIC WAY OF BUILDING REGRESSION MODELS WHEN DEALING WITH BIG DATA. Big data isn't just big. It also may come with problems, such as categories pretending to be numerical and missing data. To overcome these problems and exploit all of that data, you need to turn business insights into a statistical model. That's the task for this chapter: combining business knowledge and regression analysis into a powerful approach for successful modeling.

[1] For more examples, see the special section in the March 11, 2013 issue of the *Wall Street Journal*.

1 | MODELING PROCESS

This chapter describes how to build a regression from a large data table. We begin with an overview of the steps, organized within the framework of a 4M exercise. Most steps involve familiar methods, such as using plots to check for outliers. Other steps introduce new concepts, such as overfitting and validation, that become relevant when modeling large amounts of data. After introducing the modeling process, we illustrate the process by building a model designed to help a finance company identify profitable customers. Its clients are small businesses that purchase financial services. These services include accounting, banking, payroll, and taxes. Its clients know their market niche, but not how much to withhold in payroll taxes or the best way to negotiate the terms of a loan. The finance company also occasionally provides short-term loans to clients.

The goal of the regression model is to estimate the value of fees generated by potential clients. Clients sign an annual contract, and the finance company would like to be able to predict the total fees that will accumulate over the coming year. It currently relies on search firms, called originators, that recruit clients. The finance company would like to identify clients that will generate greater fees so that it can encourage originators to target them.

The data available for building this model describe 3,272 clients that bought services from this financial provider for the first time last year. The data include 53 variables. This is a relatively small data set compared to those at WalMart, but substantially larger and more complex than others we have used.

Modeling Steps

To build a model from a large data set, you need to be knowledgeable about the business and have a plan for how to proceed. The approach described in this chapter expands the 4M strategy used in other chapters, emphasizing issues that arise with big data. You don't have to follow this script, but experience has taught us that these steps—such as looking at histograms of data—lead to insights that save time later.

Motivation

- *Understand the context*. Every 4M analysis begins with Motivation, and this emphasis on business knowledge becomes more essential when you are dealing with big data. From the start, make sure you know what you're trying to do. You should not have to browse the data table to figure out which variable is the response. In this example, we know that we need to build a model that predicts the fees generated by clients. You also will find it helpful to set some goals for your model, such as how precisely the model should predict the response.

Method

- *Anticipate features of the model*. Which variables do you expect to play an important role in your regression? It is often smart to answer this question before looking through a list of the available variables. Once you see this list, you may not realize an important characteristic is missing. Anticipating the relevant explanatory variables also forces you to learn about the business problem that provides the context for the statistical model. In this case, that implies learning more about how a financial advisor operates.
- *Evaluate the data*. It is always better to question the quality of your data before the analysis begins than when you are preparing a summary. There's no sense spending hours building a regression only to realize that the data do not represent the appropriate population or that the variables do not measure what you thought they did. For this example, we need to be sure that the data supplied by the financial provider—observations describing the fees generated by past customers—is representative of those of

future customers it hopes to predict. The concepts of sampling introduced in Chapter 13 are relevant here.

Mechanics

- *Scan marginal distributions*. Before fitting models, browse histograms and bar charts of the available variables. By looking at the distribution of each variable, you become familiar with the data, noticing features such as measurement scales, skewness, and outliers or other rare events. If data are time series, then plots over time should be done before considering histograms.
- *Fit an initial model*. The choice of explanatory variables to include in an initial regression comes from a mix of substantive insight and data analysis. Substantive insights require knowledge of the business context along with your intuition for which features affect the response. Start with the model suggested by your understanding of the problem; then revise it as needed. Notice that we describe this model as the "initial regression." Building a regression in practice is usually an iterative process, and few stop with the first attempt.
- *Evaluate and improve the model*. Statistical tests dominate this stage of the process, and the questions are familiar:
  - Does the model explain statistically significant variation in the response, as judged by the overall *F*-statistic?
  - Do individual variables explain statistically significant variation in the response?
  - Can you interpret the estimated coefficients? Would a transformation (usually, the log of a variable) make the interpretation more sensible?
  - Does collinearity or confounding obscure the role of any explanatory variables?
  - Do diagnostic plots indicate a problem with outliers or nonlinear effects?
  - Does your model omit important explanatory variables?

tip

It is much easier to judge whether a variable in the model is statistically significant (just use the *t*-statistic) than to recognize that a variable has been left out of the model. Without understanding the substantive context, you'll never realize an essential variable is missing from the data set you're analyzing. And even if it's there, identifying an omitted variable can resemble the proverbial search for the needle in a haystack. Fortunately, statistics has tools that help you find additional variables that improve the fit of your model.

If you decide to remove explanatory variables from a regression (usually variables that are not statistically significant), be careful about collinearity. An explanatory variable may be insignificant because it is unrelated to the response or because it is redundant (collinear) with other explanatory variables. If removing variables, do it one at a time. If you remove several variables at once, collinearity among them may conceal an important effect. Variance inflation factors (Chapter 24) warn about possible collinearity.

Message

- *Summarize the results*. Save time for this step. You might have spent days building a model, but no one is going to pay attention if you can't communicate what you found using the language of the business. Illustrations that show how the model is used to predict typical cases may be helpful.

  Not everyone is going to be familiar with the details of your model, so take your time. For example, if your model includes logs, then explain them using percentages rather than leaving them a mystery. Point out if your analysis supports or contradicts common beliefs.

  Finally, no model is perfect. If there is other data that you think could make the model better, now's the time to point that out as well as your thoughts on how to improve any weaknesses.

What Do You Think?

These questions are intended to remind you of several aspects of regression analysis that you may have forgotten. We will need these soon, so it's time to scrape off the rust.

a. A regression of *Sales* (in thousands of dollars) on spending for advertisements (also in thousands of dollars) produced the fitted equation

$$\textit{Estimated}\ \text{Sales} = 47 + 25 \log_e \text{Advertising}.$$

Interpret the estimated slope. (Feel free to refer back to Chapter 20.) Describe how the gain in estimated sales produced by spending another $1,000 on advertising depends on the level of advertising?[a]

b. What plot would you recommend for checking whether an outlier has influenced a multiple regression?[b]

c. If a categorical variable has 20 levels, what problems will you encounter using it as an explanatory variable in a multiple regression?[c]

d. In a regression of productivity in retail stores, what problem will you encounter if two explanatory variables in the model are the percentage male and percentage female in the workforce?[d]

Model Validation and Overfitting

model validation Confirming that the claims of a model hold in the population.

Model validation is the process of verifying the advertised properties of a statistical model, such as the precision of its predictions or its choice of explanatory variables. In previous chapters, examples of regression name the response and explanatory variables. If you know the equation of a model, then model validation means checking the conditions, such as the similar variances condition. When we use the data to pick the relevant variables, validation requires more.

The iterative process of building a regression model requires a bit of "try this, try that, and see what works." This sort of "data snooping" typically leads to a higher R^2, but the resulting model may be deceptive. The problem arises because we let the data pick the explanatory variables. A search for explanatory variables might cause us to look at tens or even hundreds of t-statistics before finding something that appears statistically significant. The deception is that there's a good chance that a variable appears statistically significant because of data snooping rather than its association with the response. This problem is known as overfitting.

overfitting Confusing random variation for systematic properties of the population.

Overfitting implies that a statistical model confuses random variation for a characteristic of the population. A model that has been overfit to a sample won't predict new observations as well as it fits the observed sample. Suppose a regression has residuals with standard deviation $s_e = \$50$, for example. The approximate 95% prediction intervals suggest that this model ought to predict new cases (that are not extrapolations) to within about $100 with 95% probability. If the regression has been overfit to the sample, however, the prediction errors will be much larger. Prediction intervals that claim 95% coverage might include only half of the predicted values. Overfitting also leads

[a] Each 1% increase in advertising adds about $0.01(25) = \$0.25$ thousand ($250) to estimated sales. The estimated gain decreases with the level of advertising because a $1,000 increase comprises a smaller and smaller percentage change.
[b] A plot of the residuals on fitted values. (We will see a better choice later in the chapter.)
[c] Many dummy variable coefficients (19 of them) will complicate the model, and some may have large standard errors if the category has few cases.
[d] Severe collinearity (your software may crash) because the variables are perfectly redundant: %male = 100 − %female. You should use one or the other, but not both.

to spurious claims of statistical significance for individual explanatory variables. If overfit, an explanatory variable with t-statistic 2.5 might nonetheless have slope $\beta = 0$ in the population. We would come away with the impression that differences in this explanatory variable are associated with changes in the average response when they are not.

tip

There's a simple explanation for why overfitting occurs. Statistics rewards persistence! If you keep trying different explanatory variables, eventually one of them will produce a t-statistic with p-value less than 0.05. This happens not because you've found an attribute of the population, but simply because you've tried many things. We've run into this problem before. The causes of overfitting affect control charts (Chapter 14) and the analysis of variance (Chapter 26). Both procedures require many test statistics. Control charts test the stability of a process repeatedly. The analysis of variance tests for differences between the means of several groups. Comparing the means of, say, eight groups implies $(8)(7)/2 = 28$ pairwise comparisons. Because testing H_0 with $\alpha = 0.05$ produces a Type I error with probability 0.05, continued testing inevitably leads to incorrectly rejecting a null hypothesis. Incorrectly rejecting H_0: $\beta_j = 0$ in regression means thinking that the slope of X_j is nonzero when in fact X_j is unrelated to the response in the population.

The approaches taken to control overfitting in control charts or ANOVA work in regression, but these are unpopular. It's easy to see why. Both avoid the risk of too many false rejections of H_0 by decreasing the alpha level below 0.05. In control charts, for example, the default is $\alpha = 0.0027$ so that a test rejects the null hypothesis that claims the process is under control only when $|Z| > 3$. In regression, this approach would mean rejecting H_0: $\beta_j = 0$ only if the p-value of the t-statistic were less than 0.0027, or roughly if $|t| > 3$. This approach is unpopular in regression because it makes it harder to find statistically significant explanatory variables that increase R^2. Similar objections apply to the Bonferroni method in Chapter 26.

So how should you guard against overfitting?

Suppose you have hired consultants to build a regression model to help you understand a key business process, such as identifying profitable customers. How would you decide whether to believe the regression model they have developed? Maybe the consultants snooped around until they found characteristics that inflated R^2 to an "impressive level" to convince you to pay them a large fee for the model. Before you pay them and incorporate their model into your business, what would you do to test it?

test sample Data reserved for testing a model.

training sample Data used to build a model.

A popular, commonsense approach is to reserve a so-called holdback or **test sample** that the consultants do not see. The data used for fitting the model is called the **training sample**. The idea is to save data that can be used to test the model by comparing its fit to the test sample to its fit to the training sample. We can compare the fit of the model in the two samples in several ways. The most common uses the estimated model built from the training sample to predict the response in the test sample. The estimated model should be able to predict cases in the test sample as accurately as it predicts those in the training sample. A further comparison contrasts the estimated coefficients of the explanatory variables when the model is fit in the two samples. The coefficients ought to be similar in the two samples.

The idea of reserving a test sample is straightforward, but deciding how much to hold back is not obvious. Ideally, both samples should be "large." The training sample needs to be large enough to identify a good model, and the test sample needs to be large enough to validate its fit to new data. A test sample with only 50 observations lacks enough information for serious model validation unless the model is very simple. With more than 3,000 cases from the financial provider, we have more than enough for validation. We'll reserve about one-third for validation and use the rest to build the model.

What Do You Think?

a. Imagine fitting a multiple regression with a normally distributed response and ten explanatory variables that are independently randomly generated ("random noise"). What is the probability that at least one random predictor appears statistically significant?[a]
b. One test of overfitting compares how many predicted observations in the test sample lie within the 95% prediction intervals of a model. If the test sample has 100 cases, would it indicate overfitting if 80 or fewer fell inside the 95% prediction limits? Explain.[b]
c. What if the validation sample has ten cases? That is, would finding eight or fewer inside the 95% bounds suggest overfitting?[c]
d. The answers to (b) and (c) are very different. Explain the relevance for model validation.[d]

2 | BUILDING THE MODEL

This section presents a long 4M example that follows the modeling process just described to build a regression model. We already introduced the context, and the following discussion of the variables in the data adds to that introduction. Marginal notes call out key components of the steps.

Setting Goals

Two goals for the model are to predict fees produced by new clients and to identify characteristics associated with higher fees. The model should fit well—attain a reasonable R^2—and have an equation we can interpret. These goals conflict if the model requires collinear explanatory variables to produce precise predictions. Collinearity among many explanatory variables makes it hard to interpret individual partial coefficients. It is difficult to explain what it means for one variable to change "holding the others fixed" if that never happens in the data.

Use a histogram to see the variation in the response and set goals for model precision.

What does it mean to have a "reasonable R^2"? To find an answer, look at the variation in the response. The histogram of annual fees in Figure 1 shows that clients paid amounts from around \$500 to more than \$9,000, with mean \$3,100 and standard deviation \$800. The bell-shaped distribution implies that we can predict the fees for a new client to within about \$1,600 with 95% probability without regression—just use the mean plus or minus two standard deviations. It's easy to think that regression should do much better, but it is harder than it seems. For a regression to obtain a margin of error of \$500 (predict within $\pm$\$500 with 95% probability), for example, requires that R^2 approach 90%. Such a good fit rarely happens and may be unattainable. If $R^2 \approx 60\%$, then the margin of error is about \$1,000. If that's not good enough, we are left to hope that the data contain a very good predictor.

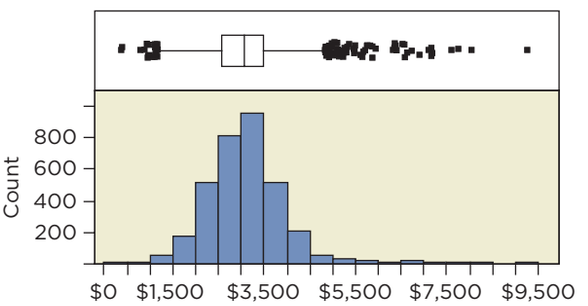


FIGURE 1 Histogram and boxplot of the annual fees paid by clients.

[a]Treating these as independent, the probability of no Type I error is $1 - 0.95^{10} \approx 0.40$.
[b]Yes, this would be very rare if the model were correct. 80 lies $(80 - 95)/\sqrt{(100(0.95)(0.05))} \approx 6.9$ SDs below the mean, a very extreme deviation. (Note: The variance of a binomial is $n\,p\,(1 - p)$.)
[c]8 lies $(8 - 9.5)/\text{sqrt}(10(0.95)(0.05)) \approx 2.2$ SDs below the mean: surprising but not shocking. The exact probability is $1 - 0.95^{10} - 10(0.95^9)(0.05) = 0.0861$—neither rare nor convincing of overfitting.
[d]The test sample in (c) is too small to detect the problem caused by overfitting.

Anticipate Features of the Model

Before we look at the list of possible explanatory variables, what would you expect to find in a regression model that predicts the value of fees? We don't work in that industry, so we do not know the ins and outs of the business. Even so, we ought to be able to list a few possible variables. For example, the size of the business seems like a good guess: clients with little revenue can't afford to pay much either, nor do they need much financial advice. The type of business that the client is in might be important, too. Those with many employees might need help with payroll. As it turns out, the data set identifies the type of the business, but does not reveal much about the size of the business outside its use of credit.

Table 1 lists the available variables and gives a short description of each. The table also shows the number of levels in categorical variables.

TABLE 1 Variables measuring the characteristics of past clients.

Variable Name	Levels	Description
Annual fees		The response, in dollars
Originator	3	Three originators recruit clients
Years in Business		Length of time the client has operated
Credit Score		Credit score of the listed owner of the business
SIC Code	270	Standard Industry Classification code
Industry Category	32	Textual description of the business
Industry Subcategory	268	More specific description of the business
Property Ownership	3	Whether the business property is owned, leased, or under mortgage
Num Inquiries		Number of times in last six months the business has sought more credit
Active Credit Lines		Number of active credit lines, such as credit cards
Active Credit Available		Total over all lines, in dollars
Active Credit Balance		Total over all lines, in dollars
Past Due Total Balance		Total over all lines, in dollars
Avg Monthly Payment		Total over all lines, in dollars
Active #30 Days		Number of times payment on a credit line was 30 days past due
Active #60 Days		Number of times payment was 60 days past due
Active #90 Days		Number of times payment was 90 days past due
Number Charged Off		Number of credit lines entering collection, total
Credit Bal of Charge-Offs		Total balance over all lines at time of collection, in dollars
Past Due of Charge-Offs		Total over all lines, in dollars
Percent Active Current		Percentage of active credit lines that have current payments
Active Percent Revolving		Percentage of credit lines that are revolving (for example, credit cards)
Num of Bankruptcies		Number of times the company has entered bankruptcy
Num of Child Supports		Number of times owner has been sued for child support
Num of Closed Judgments		Number of times a closed judgment has been entered against owner
Num of Closed Tax Liens		Number of times a tax lien has been entered against the owner

TABLE 1 Continued

Variable Name	Levels	Description
Population of ZIP Code		Refer to the ZIP Code of the business
Households in ZIP Code		
White Population		
Black Population		
Hispanic Population		
Asian Population		
Hawaiian Population		
Indian Population		
Male Population		
Female Population		
Persons per Household		
Median Age		
Median Age Male		
Median Age Female		
Employment		
First-Quarter Payroll		
Annual Payroll		
State	51	
Time Zone	6	Number of time zone, relative to Greenwich
Metro. Stat. Area	231	Name of the metropolitan area in which business is located
Region	4	Name of the region of the country
Division	9	Census division

Skim the list of potential variables. Does it contain what you expected? Do you know what the variables measure?

A quick scan of Table 1 reveals that many of these variables describe the creditworthiness of clients. There are two explanations for this. First, data that describe how a business uses credit are readily available from credit bureaus. For a fee, the financial provider can purchase credit scores, balance information, and other financial details from one of the three big credit bureaus (Experian, Equifax, and TransUnion).

The second explanation for the prevalence of these variables is more revealing: this provider often deals with businesses that have a troubled history. The variables provide detail on past problems, including a history of credit default and bankruptcy. A history of problems does not imply that a client was out to defraud others, but such a past suggests that the client has a volatile business, one with large swings from profitability to losses. As it turns out, volatile businesses are often the most profitable targets for financial service providers. These businesses have problems managing money and hence often seek—and pay for—advice.

Without insight into how such companies operate, it is hard to build a model from so many explanatory variables. Rather than rely on an automatic search (such as the stepwise algorithm described in SIA 8), let's imagine that we have a conversation with managers of the financial provider that provides us with insights as to which factors predict the fees generated by a client. (Exercise 50 exploits further comments along these lines.)

1. Clients that have a wider range of opportunities for borrowing—namely, those with more lines of credit—generate higher fees.
2. Clients that have nearly exhausted their available credit also generate higher fees.
3. Clients owned by individuals with high credit scores often have novel financial situations that produce large fees. Credit scores run from about 200 to 800, mimicking the range used in SATs. Individuals with high scores tend to be wealthy with high-paying jobs and a clean credit record.
4. Three originators recruit clients for the financial company. Managers suspect that some of the originators produce better (more profitable) clients.
5. Retail businesses in low-margin activities (groceries, department stores) operate a very lean financial structure and generate few fees. Businesses that provide services to other businesses often have complex financing arrangements that generate larger fees.
6. The financial provider handles clients with a history of bankruptcy differently. It structures the fees charged to firms with a prior bankruptcy differently from those charged to other firms.

Use the context to anticipate direction of effects and nonlinearity.

Now that we have a sense of possible explanatory variables, we can guess the direction of their effects on the response. Finding that the coefficient of an explanatory variable has the "wrong" sign may indicate that the model omits an important confounding variable. For instance, the third comment suggests that higher credit scores should come with higher fees—positive association. Finding a negative coefficient for this variable would indicate the presence of a confounding effect; for example, businesses whose owners have high credit scores might also be older (and perhaps older firms generate less fees).

Once we have the direction, we can anticipate whether the association is linear or nonlinear. Nonlinear patterns should be expected, for instance, in situations in which percentage effects seem more natural. For instance, we often expect the effects of an explanatory variable on the response to diminish as the explanatory variable grows. Do you expect the same increase in fees when comparing clients with 1 or 2 lines of credit as when comparing those with 20 or 21 lines of credit? You might, but it seems plausible that a log transformation might be helpful. The jump from 1 to 2 doubles the number of lines of credit, whereas a change from 20 to 21 is only a 5% increase. Both add one more line of credit, but the second difference (from 20 to 21) seems less important. (Refer to Chapter 20 for a detailed treatment of nonlinear association.)

Evaluate the Data

It is a given that our data must be a representative sample from the population of interest. Here we give two less obvious reasons for why data may not be representative. One is generic in business applications of statistics, and another comes from a comment from the discussion with managers.

The sixth comment listed previously points out that the finance company has a very different fee structure for clients with a history of bankruptcy. A quick peek at the data reveals a mixture of firms with and without a prior bankruptcy. Of the 3,272 cases, 378 clients (11.6%) have a bankruptcy and 2,894 do not. Taken together, these observations are a sample of the population of clients, but according to managers, the two types of clients produce very different fees.

Are the data a sample from one population? Segmentation may be needed.

We have two ways to deal with this heterogeneity. One approach keeps the data together. This approach requires a regression with numerous interactions that allow different slopes for modeling the effect on fees for firms with and without a bankruptcy. With so many possible variables, the use of interactions seems daunting; so we will take a different path. Because of the heterogeneity produced by bankruptcy, we will set aside the bankrupt clients and use only those without a bankruptcy. We can think of this subset as a sample from the population of clients that do not have a history of bankruptcy. (Exercise 52 models clients with a history of bankruptcy.)

segmentation Separating data into homogeneous subsets that share a collection of explanatory variables.

This sort of **segmentation**, dividing data into homogeneous subsets, is common when modeling big data. Part of the rationale for modeling these subsets separately is that we suspect that variables that are relevant for bankrupt businesses are not relevant for others. For instance, it might be important to know how long ago the bankruptcy occurred and use this in a model. That variable is not defined for firms without bankruptcy. (Credit bureaus routinely partition individuals into six or more segments when assigning credit scores; those with bankruptcies or little history are handled separately from others.)

After segmentation, there is another reason that data may not be representative: the passage of time. Think about how long it takes to assemble a data set such as this one. The response is the level of fees produced over a year. Hence, it takes at least a year just to observe the response, in addition to time to accumulate the observations. Let's assume that it took the company 18 months to build this data table. Now the company must build a model, instruct its staff in the use of the model, and then use that model to predict fees. How long will that take? You can see that it might easily be two years after data collection begins before the model is routinely used, and then it will take another year to see how well its predictions turn out. Three years is a long time in the business world, and much can change.

A regression model that describes the association between customer characteristics and fees is likely to be sensitive to the overall economy. The nature of customers might, for instance, change if the economy were to suddenly head down or boom. We need to confirm that our data in this example comes from a period in which the economy is similar to the economy when the model is used. The use of a test sample to evaluate the model only partly addresses this problem. The cases in the test sample come from the same collection as those used to fit the model, not from cases several years down the road. Performance in a test sample does not guarantee that a model will work well in the field if the economic environment changes.

tip

Scan Marginal Distributions

To illustrate the benefits of skimming marginal distributions, consider the histograms of two numerical and two categorical variables from the financial data in Figure 2. From this point on, we will use the data for the 2,894 clients without a bankruptcy.

Use histograms and bar charts to find outliers, skewness, data entry errors, and other anomalies.

The variable *Active Credit Lines* is right-skewed, and this type of skewness often anticipates the use of a log transformation to capture diminishing fees per additional line of credit. A log scale for this variable would produce a more symmetric distribution as well. The histogram of *Credit Score* reveals another common anomaly: missing data that are not identified as missing. The boxplot highlights a clump of three outliers with credit score 0; most likely, these are missing data that was entered as 0 because nothing was known. Actual credit scores do not extend below 200. We exclude these from further analysis.

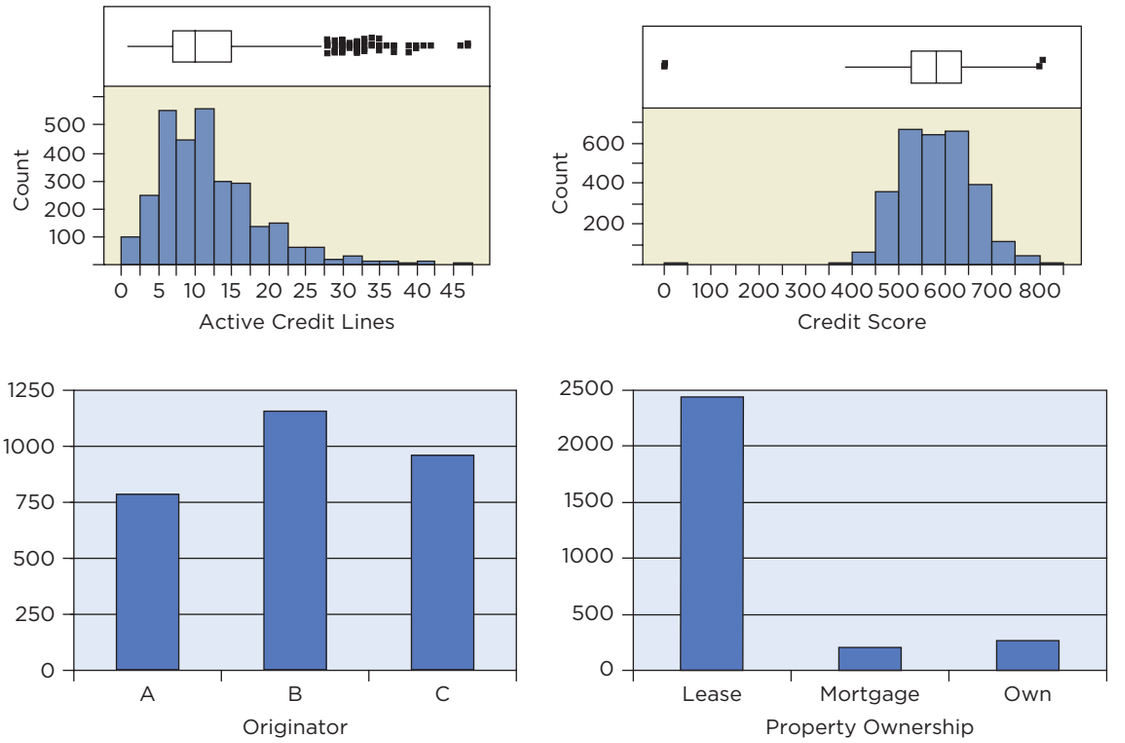


FIGURE 2 Marginal distributions of four variables show skewness, outliers, and proportions.

The bar charts of *Originator* and *Property Ownership* in Figure 2 illustrate good and bad properties of categorical variables. The bar chart of *Originator* spreads numerous observations over each of the three categories. This variable has enough information to distinguish the three groups. On the other hand, the bar chart of *Property Ownership* has few cases in two categories. We know very little about businesses that have mortgages or own their property. When fitting models, we may not have enough data to find statistically significant differences.

Combine nearly empty categories unless they have very different responses.

This problem—not enough data in some categories—gets worse as the number of categories increases. The fifth comment on page 9 suggests that the industry category affects fees. Table 2 shows the frequency distribution over the 29 categories of this variable (after removing bankruptcies and the three observations with missing credit scores). It's a good thing we checked this distribution: seven of the categories have only one observation. Estimating the effect on fees of belonging to one of these groups would mean relying on just one observation. We either must exclude these rare categories or lump them together into a category called "Other." Before combining rare groups, make sure the responses in these groups are similar (such as by using comparison side-by-side boxplots and the ideas from Chapter 26). If, for example, fees in a rare category average several times fees elsewhere, we would not want to mix that category with cases from typical categories. The means and standard deviations in Table 2 don't show any anomalous categories (which can be confirmed with plots). For this analysis, we will merge the categories with 10 or fewer cases (shaded in Table 2) into a group called "Other." The resulting categorical variable *Merged Industry* has 21 levels, all with at least 14 cases. We would probably merge more categories, but the fifth comment singles out retailers as different; so we will keep the remaining categories distinct.

TABLE 2 Frequency distribution for the variable Industry Category, with mean and standard deviation of fees in that group.

Industry	Count	Fees	
		Mean	SD
Manufacturers—Chemicals	1	$3,170	
Manufacturers—Industrial and Commercial Machinery	1	$2,830	
Manufacturers—Others	1	$3,030	
Manufacturers—Stone, Clay, Glass, and Concrete Products	1	$2,810	
Retail—Automotive Dealers & Gas Stations	76	$2,801	$740
Retail—Clothes & Accessories	157	$3,087	$625
Retail—Food Stores	166	$2,668	$687
Retail—Furniture, Furnishings, & Appliances	186	$3,010	$674
Retail—General Merchandise Stores	18	$2,718	$562
Retail—Hardware & Home Improvement	41	$2,925	$595
Retail—Others	361	$3,024	$683
Services—Amusement & Recreation	113	$3,079	$738
Services—Auto Repair & Maintenance	305	$2,974	$698
Services—Communication	1	$3,030	
Services—Construction Contractors	14	$2,939	$832
Services—Eating and Drinking	770	$3,034	$684
Services—Educational	10	$2,980	$675
Services—Electric Gas and Sanitary	1	$3,130	
Services—Farming	4	$3,173	$598
Services—For Businesses	110	$3,237	$742
Services—Health	58	$3,265	$720
Services—Home Improvement	37	$2,929	$585
Services—Hotels Rooming and Other Lodging	18	$3,146	$852
Services—Miscellaneous Repair	21	$3,175	$625
Services—Other Transportation	1	$2,110	
Services—Personal	275	$3,056	$706
Services—Pets	27	$2,975	$717
Services—Printing Publishing and Allied Industries	6	$3,242	$567
Services—Real Estate	3	$3,170	$766
Services—Social Services	8	$2,529	$654
Services—Transportation & Storage	72	$3,128	$739
Wholesale	28	$2,879	$706

Prepare for Validation

For the test sample to be useful for model validation, we must set these cases aside before fitting regression equations. The data table includes the column *Role* with values "Train" and "Test" for this purpose. We decided to use 2,000 randomly chosen observations for fitting models (about two-thirds of the data) and to reserve the remaining 891 for testing the fit (Figure 3). It is common to use more data for estimation than for testing, and one often sees 80 or 90% of the data used for fitting the model. The larger the training

Exclude the test sample before fitting models.

sample used for estimation, the more precise the parameter estimates become because standard errors decrease with increasing sample size. For this example, we keep a large testing sample to obtain precise estimates of how well the model performs.

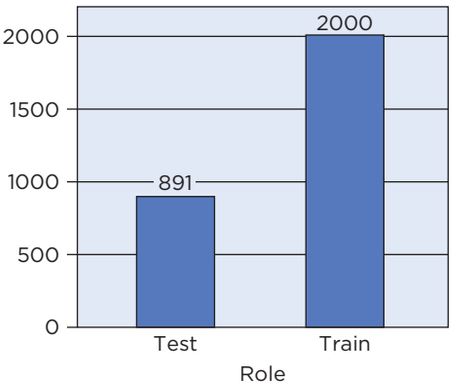


FIGURE 3 Sizes of the training and test samples.

Fit an Initial Model

Our initial model reflects what we know about context of the problem gleaned from the previous substantive comments listed on page 9. The first comment regarding the number of credit lines seems clear: businesses with more credit lines generate higher fees. So we will use *Active Credit Lines* in our model. The second comment requires that we build a variable to capture the notion that firms that have used up more of their available credit generate higher fees. To capture this effect, we define the variable *Utilization*, which is the ratio of the active credit available divided by the sum of credit available plus the balance (times 100 to give a percentage). tip Even though this data table has numerous columns, we often must construct yet more variables to incorporate the desired features into a regression. The third comment suggests that the model should include *Credit Score* as another explanatory variable.

Use a scatterplot matrix to see the relationships among the response and explanatory variables.

Figure 4 shows the scatterplot matrix for the response and these explanatory variables, color-coded to identify the originator noted in the fourth comment on page 9. Red denotes businesses found by Originator A, green denotes B, and blue denotes C. Color-coding works well for *Originator* because it has only three categories. tip Color-coding is helpful for categorical variables with few levels but confusing for variables with many levels, such as the industry of these companies.

None of these variables has a very high correlation with the response, *Annual Fees*. The largest correlation is with *Active Credit Lines* ($r = 0.33$). There is little correlation with *Utilization* ($r = 0.15$) and a surprising small negative correlation with *Credit Score* ($r = -0.10$). Little collinearity connects these three explanatory variables; the largest correlation among them is the negative correlation between *Utilization* and *Credit Score* ($r = -0.22$, which could be expected since one's credit score goes down as more of the available credit is used). Don't be disappointed by the small correlations of these explanatory variables with the response. We're just starting and need to combine them with the other effects in a multiple regression. Although the numerical variables have small correlations with the response, this plot shows a large effect for *Originator*. The elevated positions of blue points in the top row of the scatterplot matrix indicate higher fees coming from businesses found by Originator C.

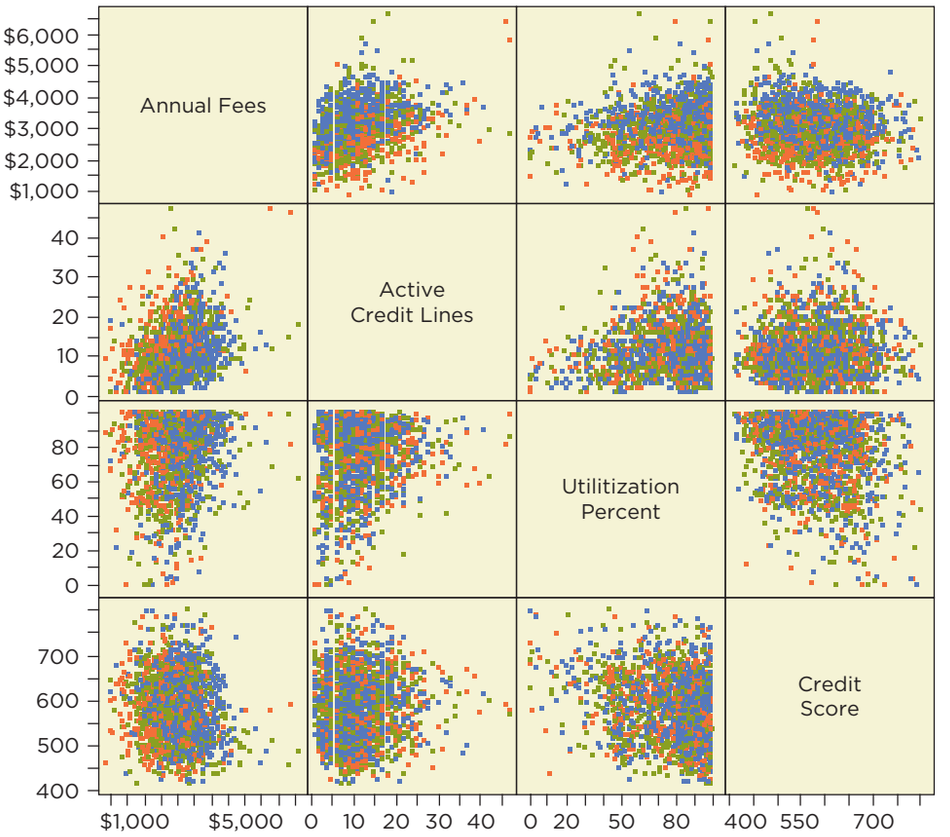


FIGURE 4 Scatterplot matrix of the response and several potential explanatory variables, color-coded by the originator for each client.

Choose an initial model based on substantive insights.

We fit a multiple regression with these variables, along with the response, *Merged Industry*. The output in Table 3 summarizes the results. Overall, the model is highly statistically significant. Although it is premature to believe that the Multiple Regression Model holds for this equation (we have not added all of the relevant explanatory variables so that nonrandom structure remains in the residuals), the overall F-statistic ($F = 26, p < 0.0001$) shows that one is unlikely to obtain $R^2 = 0.25$ with $n = 2{,}000$ and $k = 25$ explanatory variables by chance alone. Some of these explanatory variables are related to the response. Light shading in the table of estimates highlights coefficients that are statistically significant ($p < 0.05$).

R^2	0.2476
s_e	610.9011
n	2000

TABLE 3 Initial multiple regression for client fees.

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	25	242423024	9696921	25.9832
Error	1974	736697079	373200	**Prob > F**
C. Total	1999	979120104		<0.0001*

Parameter Estimates				
Term	**Estimate**	**Std Error**	***t*-statistic**	***p*-value**
Intercept	2935.16	210.57	13.94	<0.0001
Active Credit Lines	36.92	2.17	17.01	<0.0001
Utilization Percent	3.86	0.72	5.40	<0.0001
Credit Score	−1.05	0.20	−5.36	<0.0001
Originator [A]	−541.05	35.47	−15.25	<0.0001
Originator [B]	−259.24	32.80	−7.90	<0.0001
Merged Industry [Other]	9.70	202.95	0.05	0.9619
Merged Industry [Retail—Automotive Dealers & Gas Stations]	3.83	175.36	0.02	0.9826
Merged Industry [Retail—Clothes & Accessories]	335.94	164.12	2.05	0.0408
Merged Industry [Retail—Food Stores]	−77.23	163.37	−0.47	0.6365
Merged Industry [Retail—Furniture, Furnishings, & Appliances]	256.09	161.82	1.58	0.1137
Merged Industry [Retail—General Merchandise Stores]	−108.86	246.59	−0.44	0.6589
Merged Industry [Retail—Hardware & Home Improvement]	88.70	191.98	0.46	0.6441
Merged Industry [Retail—Others]	199.04	158.16	1.26	0.2084
Merged Industry [Services—Amusement & Recreation]	184.38	168.40	1.09	0.2737
Merged Industry [Services—Auto Repair & Maintenance]	223.36	159.22	1.40	0.1608
Merged Industry [Services—Construction Contractors]	36.74	255.22	0.14	0.8856
Merged Industry [Services—Eating and Drinking]	244.06	155.79	1.57	0.1174
Merged Industry [Services—For Businesses]	384.72	167.76	2.29	0.0219
Merged Industry [Services—Health]	341.63	179.85	1.90	0.0576
Merged Industry [Services—Home Improvement]	36.62	203.53	0.18	0.8572
Merged Industry [Services—Hotels Rooming and Lodging]	61.78	246.91	0.25	0.8024
Merged Industry [Services—Miscellaneous Repair]	382.62	224.02	1.71	0.0878
Merged Industry [Services—Personal]	221.21	159.33	1.39	0.1652
Merged Industry [Services—Pets]	−26.88	213.38	−0.13	0.8998
Merged Industry [Services—Transportation & Storage]	247.09	179.44	1.38	0.1687

Check signs of coefficients, being mindful of collinearity and confounding due to omitted variables.

Most of the estimated coefficients that are statistically significant in Table 3 have the expected sign. Adding more credit lines adds to fees (about $37 per additional line), as does having a higher utilization (although only $4 per additional 1% increase). *Credit Score*, however, is statistically significantly negative. This unexpected sign suggests confounding due to an omitted variable; otherwise, the managers who made suggestions on page 9 were wrong.

The diagnostic plot of the residuals on fitted values in Figure 5 looks okay, although the coloring might be confusing. This plot is colored as in Figure 4 by the three originators. Notice that the red points (Originator A) tend to lie to the left side of the plot, with the blue points (C) dominating the right side and the green points (B) in the middle. This pattern does not indicate a problem and instead confirms the effect of *Originator*. Red points fall on the left-hand side of the plot because clients obtained by this originator produce systematically lower estimated fees than those from the other two. How much lower? The estimated coefficient for the dummy variable *Originator[A]* in Table 3 shows that if comparing fees from otherwise comparable clients, fees for business identified by Originator A average about $540 less than those from Originator C. Similarly, the green points for Originator B lie in the middle of the

plot because fees from its clients average about $260 less than those from Originator C.

FIGURE 5 Scatterplot of residuals on fitted values for the initial multiple regression model, colored by the three originators.

Evaluate the Model

We need to be careful when thinking about the importance of *Merged Industry* in this model. First, recall that the coefficients of dummy variables that represent the levels of a categorical variable compare the average response in the category shown to the average response in a baseline category, the category not represented by a dummy variable. (If necessary, review these ideas in Chapter 25.) Which industry defines the baseline? Our software automatically generates dummy variables and by default leaves out the last one alphabetically, which happens to represent wholesalers. Hence, all of the coefficients in Table 3 for *Merged Industry* are comparisons to wholesalers. If you change the baseline, then all of these estimates change. Also, if the baseline category has but a few cases, then many of the estimated comparisons to the baseline will not be statistically significant because of the small sample size in the baseline category. Here 16 of the 2,000 businesses in the training sample are wholesalers. Consequently, it is often a good idea to use a category with many cases to define the baseline. For example, 523 businesses in the training sample are restaurants (with the label "Services—Eating and Drinking"). Using this category to define the baseline produces the fitted estimates shown in Table 4.

tip

R^2	0.2476
s_e	610.9011
n	2000

TABLE 4 Estimated initial multiple regression for client fees with the baseline industry category changed to be restaurants rather than wholesalers.

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Ratio**
Model	25	242423024	9696921	25.9832
Error	1974	736697079	373200	**Prob > F**
C. Total	1999	979120104		<0.0001*

Parameter Estimates				
Term	**Estimate**	**Std Error**	***t*-statistic**	***p*-value**
Intercept	3179.23	143.38	22.17	<0.0001*
Active Credit Lines	36.92	2.17	17.01	<0.0001
Utilization Percent	3.86	0.72	5.40	<0.0001
Credit Score	−1.05	0.20	−5.36	<0.0001
Originator [A]	−541.05	35.47	−15.25	<0.0001
Originator [B]	−259.24	32.80	−7.90	<0.0001
Merged Industry [Other]	−234.36	136.35	−1.72	0.0858
Merged Industry [Retail—Automotive Dealers & Gas Stations]	−240.23	89.03	−2.70	0.0070*
Merged Industry [Retail—Clothes & Accessories]	91.88	63.25	1.45	0.1465
Merged Industry [Retail—Food Stores]	−321.29	62.30	−5.16	<0.0001*
Merged Industry [Retail—Furniture, Furnishings, & Appliances]	12.03	58.25	0.21	0.8364
Merged Industry [Retail—General Merchandise Stores]	−352.92	195.14	−1.81	0.0707
Merged Industry [Retail—Hardware & Home Improvement]	−155.37	118.63	−1.31	0.1905
Merged Industry [Retail—Others]	−45.03	46.55	−0.97	0.3335
Merged Industry [Services—Amusement & Recreation]	−59.68	74.15	−0.80	0.4210
Merged Industry [Services—Auto Repair & Maintenance]	−20.70	49.88	−0.42	0.6782
Merged Industry [Services—Construction Contractors]	−207.32	205.62	−1.01	0.3134
Merged Industry [Services—For Businesses]	140.66	73.59	1.91	0.0561
Merged Industry [Services—Health]	97.57	98.30	0.99	0.3210
Merged Industry [Services—Home Improvement]	−207.44	136.01	−1.53	0.1274
Merged Industry [Services—Hotels Rooming and Lodging]	−182.28	195.28	−0.93	0.3507
Merged Industry [Services—Miscellaneous Repair]	138.56	165.54	0.84	0.4027
Merged Industry [Services—Personal]	−22.85	51.43	−0.44	0.6569
Merged Industry [Services—Pets]	−270.95	150.75	−1.80	0.0724
Merged Industry [Services—Transportation & Storage]	3.03	97.07	0.03	0.9751
Merged Industry [Wholesale]	−244.06	155.79	−1.57	0.1174

Changing the baseline category only affects the intercept and the coefficients associated with the categorical variable. The intercept now refers to restaurants rather than wholesalers, and the coefficients of the other dummy variables for *Merged Industry* make comparisons to restaurants. As a result, these estimates decrease by 244.06, the coefficient of the wholesale dummy variable in Table 4. That shi ft pushes some estimates closer to zero and others farther away. Also, with a larger baseline category (523 restaurants versus 16 wholesalers), the estimates have smaller standard errors. For example, the standard error of the coefficient of the dummy variable identifying retail food stores is 163 in the initial regression with wholesalers as the baseline but 62 in this fit. The estimated coefficient is also more negative, implying that fees for retail food stores are statistically significantly less than for restaurants (given the other variables in the model). The overall summary of the model ($R^2 = 0.2476$ with $F = 25.98$) and the coefficients for the other explanatory variables remain the same. For example, the slope for the number of active credit lines remains 36.92 and that for utilization remains 3.86.

When a categorical variable has just a few levels, such as *Originator*, it is not too hard to sort out the consequences of picking a baseline category. The presence of many categories, however, becomes confusing. It is easy to inadvertently choose a nearly empty baseline category that produces no statistically significant coefficients even though there are large differences among the categories. To avoid that problem and obtain a better test of the importance of a categorical variable, use the partial *F*-statistic.

partial *F*-statistic Used to test H_0 that a subset of coefficients in a regression all have coefficient 0.

A **partial *F*-statistic** tests whether a subset of coefficients in a regression model simultaneously have coefficient zero. You can think of this statistic as testing the improvement in R^2 obtained when adding several variables as a group to a regression, as happens when we add a categorical variable with several levels. This statistic compares the fit obtained by two regressions, a regression that includes the full set of explanatory variables, and a partial regression that removes one or more explanatory variables (thereby setting their coefficients to zero). The test answers the question: Does the full regression explain statistically significantly more variation in the response than the partial regression?

The formula for the partial *F*-statistic is:

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{partial}})/(k_{\text{full}} - k_{\text{partial}})}{(1 - R^2_{\text{full}})/(n - k_{\text{full}} - 1)} \quad \frac{(\text{Change in } R^2)/(\text{Number of Added Terms})}{(\text{Remaining Variation})/(\text{Residual d.f.})}$$

$$= \frac{R^2_{\text{full}} - R^2_{\text{partial}}}{1 - R^2_{\text{full}}} \times \frac{n - k_{\text{full}} - 1}{k_{\text{full}} - k_{\text{partial}}}$$

The partial *F*-statistic compares the change in R^2 per added variable to the amount of unexplained variation per residual degree of freedom. In this formula, the subscript "full" identifies properties of the regression model using all of the variables, and the subscript "partial" identifies the regression that sets a subset of the coefficients to 0. The constant k refers to the number of estimated coefficients (not including the intercept). The second way of writing the formula stresses the importance of the sample size n in determining the size of the *F*-statistic.

Use the partial *F*-statistic to judge the contribution of complex categorical explanatory variables.

As an example of the partial *F*-statistic, we test the null hypothesis that adding the collection of 20 dummy variables that represent *Merged Industry* does not improve the fit.

H_0: all β coefficients associated with *Merged Industry* are zero.

We have already fit the full regression and found (keep extra digits for R^2 to avoid too much round-off error)

$$R^2_{full} = 0.2476, k_{full} = 25, n = 2000.$$

If the model omits the dummy variables representing *Merged Industry*, the summary statistics are

$$R^2_{\text{partial}} = 0.2237, \quad k_{\text{partial}} = 5.$$

Plugging these into the formula for the partial *F*-statistic, we obtain

$$F = \frac{0.2476 - 0.2237}{1 - 0.2476} \times \frac{2000 - 25 - 1}{25 - 5} \approx 0.03177 \times 98.7 \approx 3.14.$$

The change in R^2 is relatively small until you multiply it by the ratio of residual degrees of freedom to the number of added variables, which is almost 100 to 1.

To determine whether the observed F-statistic is statistically significant at the chosen α level, compare it with the $100(1 - \alpha)$ percentile of the F distribution with $k_{\text{full}} - k_{\text{partial}}$ and $n - k_{\text{full}} - 1$ degrees of freedom (these are the integers in the preceding formula). The 5% critical value with $k_{\text{full}} - k_{\text{partial}} = 20$ added variables and $n - k_{\text{full}} - 1 = 1{,}974$ residual degrees of freedom in the full model is $F_{0.05,\,20,\,1974} = 1.576$. We reject H_0 because the observed $F = 3.14$ exceeds this threshold. (The p-value of the test is less than 0.0001, and the test rejects H_0 at any reasonable α level.) The test implies that adding this collection of 20 dummy variables (representing the 21 industries defined by *Merged Industry*) produces a statistically significant improvement in the fit of the model. Even though the coefficients of most of these dummy variables are not statistically significant, the addition of this collection of variables produces a statistically significantly better fit.

What Do You Think

We used a partial F-statistic to test the benefit of adding the collection of 20 dummy variables representing *Merged Industry* to this model. Why didn't we use the same procedure to test the addition of *Originator*? The answer is that we should, even though both coefficients for Originator are highly statistically significant. The R^2 without *Originator* in the model is 0.1589.

a. State the specific null hypothesis that is tested if we use the partial F-statistic to test the effect of adding *Originator* to the model.[a]
b. What is the appropriate value to use for k_{partial}?[b]
c. What is the value of the partial F-statistic?[c]
d. Does the addition of *Originator* produce a statistically significant increase in the R^2 of the reduced model? Explain.[d]

Partial Regression Plots

In addition to a rich set of categorical variables, complex regression models likely have several numerical explanatory variables as well. The more variables in the model, the harder it becomes to rely on scatterplots of Y on individual explanatory variables to check whether outliers and leverage points influence the estimated coefficients. Collinearity among the explanatory variables can produce surprising leverage points that influence the regression in ways that would not be expected from the usual scatterplots.

partial regression plot
Scatterplot that shows how data determine the partial slope of an X in multiple regression.

Partial regression plots remedy this weakness of scatterplots. A **partial regression plot** provides a "simple regression view" of the partial slope for each numerical explanatory variable. Fitting a line in the scatterplot of Y on X reveals the marginal slope of the simple regression of Y on X. This plot makes simple regression "simple." Once you see the scatterplot of Y on X, you can see where the line goes and estimate the slope. You can't do that for multiple regression because the model has several explanatory variables. That's where partial regression plots become handy. Partial regression plots do for multiple regression what the scatterplot does for simple regression. Fitting a line in the partial regression plot of Y on X produces the partial slope of the explanatory variable.

tip

[a] H_0: $\beta_A = \beta_B = 0$; the coefficients of the dummy variables for these categories are both zero.
[b] $k_{\text{partial}} = 25 - 2 = 23$ counts the number of coefficients in the reduced model after removing *Originator*.
[c] $F = (0.2476 - 0.1589)/(1 - 0.2476)*(2000 - 25 - 1)/(25 - 23) \approx 116.4$.
[d] Yes. The 5% critical point in the $F_{2,1974}$ distribution is 3.000. The observed F is much larger.

That also means we must look at several of these, one for each numerical explanatory variable. The construction of partial regression plots also reveals how the partial slope of an explanatory variable in multiple regression "controls for other variables."

Partial regression plots are most useful when the explanatory variables in a multiple regression are collinear. To illustrate this situation, we'll add several variables to the initial regression model. The variance inflation factors for the three numerical variables in the initial regression are less than 1.1; collinearity does not affect these estimates. The variables we will add to the model produce much more collinearity.

The variables *Active #30 Days*, *Active #60 Days*, and *Active #90 Days* count the number of occasions in which payments on a line of credit slipped to being one month, two months, or three months late. Missing a payment leaves a mark on your credit record, and as a result, these variables are correlated with *Credit Score*. The correlation matrix in Table 5 shows that these three additional variables are negatively correlated with *Credit Score*, as well as positively correlated with each other.

TABLE 5 Counts of late payments are negatively correlated with *Credit Score* and positively correlated with each other.

	Credit Score	Active #30Days	Active #60Days	Active #90Days
Credit Score	1.0000	−0.4107	−0.3781	−0.2610
Active #30Days	−0.4107	1.0000	0.7632	0.4595
Active #60Days	−0.3781	0.7632	1.0000	0.6758
Active #90Days	−0.2610	0.4595	0.6758	1.0000

Because of these correlations, adding these measures of risky credit practices affects the estimated slope for *Credit Score*. Table 6 summarizes the regression after these three variables are added.

R^2	0.3626
s_e	562.7

TABLE 6 Summary of the multiple regression after counts of late payments are added. (Coefficients are not shown for all industries.)

Parameter Estimates					
Term	**Estimate**	**Std Error**	***t*-statistic**	***p*-value**	**VIF**
Intercept	2206.29	143.80	15.34	<0.0001	
Active Credit Lines	35.59	2.01	17.73	<0.0001	1.0
Utilization Percent	4.36	0.66	6.60	<0.0001	1.1
Credit Score	0.36	0.20	1.82	0.0695	1.3
Originator [A]	−527.75	32.69	−16.15	<0.0001	
Originator [B]	−281.82	30.24	−9.32	<0.0001	
Merged Industry [Other]	−201.39	125.65	−1.60	0.1091	
	. . .				
Merged Industry [Wholesale]	−157.64	143.58	−1.10	0.2724	
Active #30Days	8.93	2.15	4.16	<0.0001	2.6
Active #60Days	23.06	6.52	3.54	0.0004	3.6
Active #90Days	20.82	2.74	7.59	<0.0001	1.9

Adding these three explanatory variables improves several aspects of the model. First, the overall fit is better, with R^2 growing from 0.25 to 0.36. In addition to the large t-statistics obtained by each added variable, the partial F-statistic confirms that adding the three together produces a statistically

significant improvement in the fit of this model. Second, consider the slope of *Credit Score*. It has become positive, although not quite significant at the usual 0.05 level. Now that we have accounted for late payments, the fit indicates that—given the other characteristics—businesses whose owners have higher credit scores produce larger fees, as suggested by the third comment on page 9. The slope of *Credit Score* was negative previously because high credit scores imply few late payments, and late payments produce fees. Now that the model separates the effects of late payments, these variables no longer confound the slope of *Credit Score*. Notice that these complications due to collinearity occur even though the magnitudes of the VIF statistics are small.

If variables are collinear, use partial residual plots to check for leverage points influencing coefficients in the regression.

This explanation for the positive slope of *Credit Score* suits the context, but is it right? Perhaps an outlier changed the sign of the slope of *Credit Score*. To find out, we use a partial regression plot. We will build the partial regression plot for *Credit Score* in this regression by hand, but in general, we defer to software. A partial regression plot is a scatterplot of residuals from two regressions. The two regressions have the same explanatory variables but different responses. One regression removes the effects of other explanatory variables from the response; in this case, regress *Annual Fees* on all of the explanatory variables in the model but for *Credit Score*. Save the residuals from this fit. The second regression removes the effects of the other explanatory variables from *Credit Score*. Regress *Credit Score* on the other explanatory variables and save the residuals. The partial residual plot for *Credit Score* is the scatterplot of the residuals of *Annual Fees* on the residuals of *Credit Score* shown on the right in Figure 6. The plot of the residuals of *Annual Fees* on the residuals from *Credit Score* is free of associations with the other explanatory variables, isolating the partial slope for *Credit Score*.

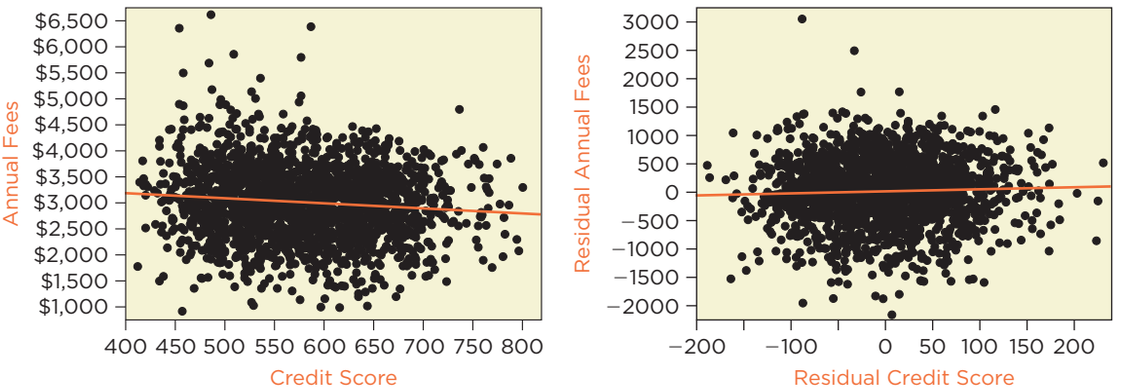


FIGURE 6 Scatterplot of Annual Fees on Credit Score and the partial regression plot for Credit Score.

Figure 6 contrasts the scatterplot of *Annual Fees* on *Credit Score* with the corresponding partial regression plot. Both show lines obtained by fitting a least squares regression to the data in each plot. The slope on the left is negative; this corresponds to the counterintuitive negative correlation between *Credit Score* and *Annual Fees*. The slope on the right is positive, matching the partial slope for *Credit Score* in the multiple regression that includes counts of late payments. The differences are subtle because *Credit Score* has a limited effect on this regression. Nonetheless, the change is important for interpretation of the model and not produced by outliers.

Figure 7 shows a more dramatic difference between a scatterplot and the corresponding partial regression plot. This figure contrasts the usual scatterplot for *Active #60 Days* with its partial regression plot. To produce the partial residual plot, we followed the same routine. First, regress the response on all of the other explanatory variables but for *Active #60 Days*. Then repeat that

regression but with *Active #60 Days* as the response. Save the residuals from both regressions; then graph the residuals of the response on the residuals of the explanatory variable *Active #60 Days*.

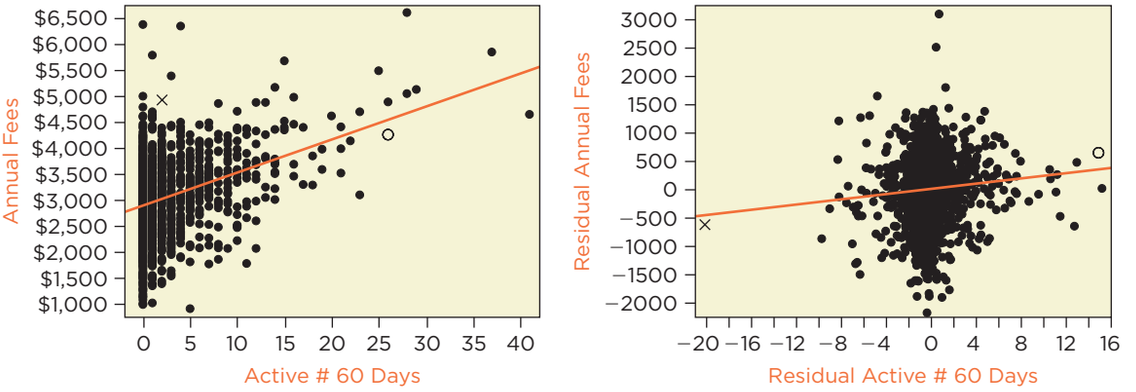


FIGURE 7 Scatterplot of Annual Fees on Active #60 Days and the Partial Regression Plot for Active #60 Days.

Collinearity explains the difference between the scatterplot and the partial regression plot. Marginally, the scatterplot on the left shows a strong effect for *Active #60 Days*. The slope in the simple regression is about $63 per credit line that was ever 60 days past due. The partial regression plot on the right shows that after accounting for the information in the other explanatory variables—namely, *Active #30 Days* and *Active #90 Days*—the estimate of the partial slope relies on leveraged points.

These leverage points do not stand out in the marginal scatterplot. Consider the leveraged outlier at the far left of the partial regression plot marked with ×. This point represents a business with about 20 fewer credit lines that went 60 days past due than expected from its other characteristics. This business also generated about $500 less in fees than expected, even though it paid large fees (close to $5,000, as shown in the scatterplot on the left). Now consider the leverage point at the right of the partial regression plot marked with ○. This business had about 15 more late events and more fees than expected.

These two points influence the model. If we were to exclude just these two out of 2,000 points, the partial slope of *Active #60 Days* would drop by 5% from 23.06 to 21.91 with a larger standard error (rising to 6.84 from 6.52). All of this happens with VIF = 3.6. The possibility for surprising outliers grows with the level of collinearity.

Improve the Model

We have only started building a model that describes the annual fees produced by the clients of the financial company. The data include other explanatory variables that may further improve the fit of this regression.

Improvements don't necessarily require adding more explanatory variables. For instance, we ought to consider whether the effects we have found should be linear or nonlinear. That choice comes down to a question of fit and interpretation. Perhaps fees do rise with every additional incident of late payment, but it might be more sensible to think in terms of percentages. Do you really think that the effect on fees of going from 1 to 2 late payments is the same as the effect of going from 20 to 21? That seems implausible, and Exercise 49 takes this notion further.

Certainly, the best way to identify more explanatory variables is to learn more about the business. Barring that, one should also take advantage of modern automatic tools.

tip

At a minimum, never end the modeling process without using an automatic search to identify other variables among those that are not substantively motivated. Well-known tools such as stepwise regression work fine, and the current literature in statistics is full of modern descendants of that procedure.

Add further variables identified by substantive insight and automated searches.

As an illustration, we use forward stepwise regression to explore the collection of available demographic variables. (Statistics in Action: Automated Modeling describes stepwise regression in greater detail, and exercises at the end of this chapter consider other financial variables.) The demographic variables describe the composition of the population, incomes, and geography where each business operates.

tip

Automatic routines that search for other explanatory variables are best used when seeded with a substantively motivated model whose coefficients appear meaningful, such as the model in Table 6. Rather than search for a model with nothing in hand, begin the search with a model that includes the explanatory variables that the context dictates. Then rely on an automatic search to find others. The automatic search will run faster and more consistently than if you were to search the other variables manually.

Table 7 summarizes the results of the stepwise search. The search began from the model in Table 6 that includes the counts of late payments. The stepwise procedure then searched the variables *Population in ZIP Code* through *Annual Payroll*, as listed in Table 1. The search excluded *State* and *Metropolitan Statistical Area* because of the huge number of categories these represent. (Exercise 46 suggests a method to use with complex categorical variables.) Setting these aside, the stepwise search considered adding 23 variables to our substantive model (treating the bundle of dummy variables for a categorical variable as "one variable").

TABLE 7 Results of a forward stepwise search of the demographic variables.

Step	Variable Added to Model	*p*-value When Entered	R^2
1	Average House Value	0.0000	0.3689
2	Median Age Male	0.0084	0.3711
3	Employment	0.0351	0.3725
4	Persons per Household	0.0415	0.3738
5	Black Population	0.2845	0.3742
6	Indian Population	0.4260	0.3744
7	Region	0.5591	0.3751

The first variable identified by stepwise search is *Average House Value*, and adding it to the model boosts R^2 from 0.3626 to 0.3689. The p-value of adding this variable is very small; consequently, adding *Average House Value* produces a statistically significant improvement at any reasonable α level. The second variable added, *Median Age Male*, produces a smaller increment in R^2, up to 0.3711 with p-value 0.0084. This p-value is less than 0.05, but should we be impressed given that we are searching through 23 variables? The Bonferroni rule (Chapter 26) suggests a smaller threshold for statistical significance when searching many terms—namely, $0.05/23 \approx 0.0022$ (the usual level divided by the number of possible terms). By that standard, *Median Age Male* is not statistically significant and should not be added to the model. This cautious use of stepwise regression stops after adding one variable, *Average House Value*.

If one were to end the modeling process here, then it would be the time to check for heteroscedasticity and normality in the residuals. For this regression, exercises show that other explanatory variables remain to be found.

Their omission leaves nonrandom structure in the residuals and produces a small deviation from normality that can be detected in the normal quantile plot. The absence of normality in the residuals of a tentative model may not be a bad thing when you are exploring data for more predictors because it may indicate that other useful explanatory variables remain to be found. With so much data, the normal quantile plot is very sensitive to even slight deviations from a normal distribution.

tip

Summary

Building a regression model from big data follows a modeling process that defines a series of stages that leverage your business knowledge. Start by setting reasonable goals for summary statistics such as R^2 and s_e based on the variation in the response. Then use what you know of the context to anticipate the explanatory variables that ought to be in the model. Managers may not understand regression analysis, but most have intuitive insights that you can convert into explanatory variables. Be mindful that some of these variables might not be in your data.

For the variables you do have, scan their marginal distributions, with an eye toward anomalies that complicate regression. These problems include outliers and unusually shaped distributions. Skewness often anticipates transformations, and multimodal distributions suggest possible segments. Recode categorical variables with many nearly empty categories to avoid cluttering a model with meaningless, insignificant estimates. If the data cannot be considered a sample from a single population, then segmentation might produce a more homogeneous sample.

Before you start fitting regressions, reserve a test sample to validate your final model later. Save enough to be able to evaluate your model, but not so much that you don't have enough data to find important explanatory variables. Then fit an initial regression to the training sample based on your intuition and what you can learn about the business context. Scatterplot matrices are helpful for seeing how the explanatory variables are related to the response and each other. Color-coding these plots is helpful for distinguishing a few categories. Be cautious if you discover collinearity among the explanatory variables. If you do, use variance inflation factors to monitor how collinearity affects your estimates and use partial regression plots to check for leverage points hidden by the collinearity.

Once you have an initial model, consider what other variables to add and which you might want to remove. Discovering that an explanatory variable has the wrong sign often indicates that the model is missing a confounding variable. Statistical tests quantify the impact of explanatory variables on the model, with t-statistics for numerical variables and partial F-statistics for multilevel categorical variables. To avoid problems with collinearity, remove variables one at a time. The coefficients of categorical variables sometimes become more meaningful if the baseline group is changed; the baseline group should represent one of the larger categories.

Before you finish, check the conditions attached to all regression models, such as the similar variances and nearly normal conditions. Deviations from these assumptions may suggest a better model. You should also consider using an automated method such as stepwise regression to explore your data for factors you might not have time to examine yourself. Use automatic methods that carefully pick variables for you because these will easily overfit the data and produce a model that cannot predict new data nearly so well as the model fits the training sample. Fortunately, before you make false promises, you can use the test sample you set aside at the start of the modeling to check for overfitting.

3 | VALIDATING THE MODEL

This section demonstrates how to use a test sample as part of the validation of a model. Validation means many things, but key among them is that predictions of the model should perform as advertised. There's a very simple test of the advertised claim: compare the standard deviation of the residuals in the training sample s_e to the standard deviation of prediction errors in the test sample. These should be close, with the test sample producing a slightly higher standard deviation. If the standard deviation of the prediction errors in the test sample is substantially larger than in the training sample, the regression has likely been overfit.

Check your model for evidence of overfitting using predictions of the test sample.

The prediction errors in the training sample are the residuals, and those in the test sample are the deviations $Y - \hat{Y}$. The underlying theory shows that if the model has not been overfit to the data, then the expected squared prediction error in the test sample should be approximately $1 + 2k/n$ times the expected squared prediction error in the training sample (see Behind the Math: Prediction Errors). This relationship implies that the sample variance in the test sample ought to be slightly larger than the residual standard deviation. If we let s^2_{test} denote the sample variance of the prediction errors in the test sample and again let k denote the number of estimated coefficients in the regression, then we should expect to find

$$s^2_{test} \approx \left(1 + \frac{2k}{n}\right) s^2_e.$$

After *Average House Value* is added, the residual standard deviation is $s_e = \$555.50$. The standard deviation of the 891 prediction errors in this test sample is slightly smaller, $s_{\text{test}} = 551.13$. Overfitting would produce larger standard deviations in the test sample. The similarity of these estimates is evidence that the estimated model has not been overfit. It predicts new data to the precision claimed by the summary of the fitted model.

Prediction intervals offer a more detailed test of a model. Rather than use the squared prediction errors, count how many observations in the test sample lie within the appropriate intervals. For example, 67% of the test sample should lie within the 67% prediction intervals (which are roughly of the form $\hat{Y} \pm s_e$) and 95% should lie within the 95% prediction intervals (roughly $\hat{Y} \pm 2s_e$).

Working with the same regression, 848 of the 891 predicted values lie within the 95% prediction intervals (95.17%). For the 67% intervals, 646 (72.50%) lie within the limits. The 95% limits are right on target. The confidence interval for the coverage of these intervals based on the test sample includes the target 0.95:

$$0.9517 \pm 1.96(0.0072) \approx 0.938 \text{ to } 0.966.$$

For the 67% intervals, the confidence interval for the coverage of the prediction intervals omits the target 0.67:

$$0.7250 \pm 0.974(0.0150) \approx 0.710 \text{ to } 0.740.$$

This is further confirmation of what we noticed at the end of the prior section: the errors from this model are not normally distributed.

You can use the test sample in a variety of other ways. For instance, you can refit the model on the test sample and see how well the estimated coefficients agree with those produced by the fit to the training sample. Estimates in the test sample ought to be similar. The key to all of these diagnostics is to have a test sample in the first place.

Best Practices

- *Spend time learning the business context of your data*. You'll be able to assemble a decent model and interpret the results. In the example of this chapter, the only reason something seemed wrong with the negative estimate of the slope of *Credit Score* was that managers said that this variable should have a positive slope.
- *Reserve a test sample for model validation*. You may not think you need it, but after spending a few days building a model, it's important to be able to show others that you've captured real structure and not just coincidental association.
- *Spend time with the marginal distributions of variables*. It is tempting to rush past this step to regression, but you're likely to get confused by variables with sparse categories, odd distributions, and data entry errors if you bypass this step.
- *Use a partial F-statistic to assess the benefits of adding a collection of variables*. If you don't, then collinearity can hide the fact that some of the added variables improve the model. This comment is particularly relevant for categorical variables. If the categorical variable has more than two levels, use the partial F-statistic to test the benefit of adding this variable to your model.
- *Check partial regression plots for leverage points*. These adjust for collinearity that can conceal the influence of leverage points in the multiple regression.
- *Use automatic methods to expand a substantively motivated initial model*. Start with the variables that the context says matter. You'll need to address them at some point when you show your model to others. Once you have milked all you can from the substantive knowledge that you have, use an automatic procedure such as stepwise regression to dredge for any remaining useful variables.
- *Use software to find p-values for the F distribution*. This distribution has two indices, one for the numerator and one for the denominator. There's no easy way to guess the threshold for statistical significance, and most textbook tables, this one included, only provide tables for small values of the numerator degrees of freedom.

Pitfalls

- *Don't assume that your data are a random sample from one population*. Describing several fundamentally different groups with a single regression requires a complex, unwieldy equation. Segmentation keeps things simpler.
- *Don't use too many segments*. Just don't get carried way with the idea of segmentation. You still need enough data in the groups to produce accurate parameter estimates.
- *Don't assume that your data include all of the relevant explanatory variables*. It can be hard to find the right explanatory variables in a large data table. It is harder to find them if they are not in the data table. Only a good understanding of the problem will help you recognize that the data table is incomplete.
- *Don't rush to fit a regression model without getting some familiarity with the data first*. Otherwise, it is far too easy to include a categorical variable with, for instance, 270 levels and poorly represented categories. Your hurrying also means that you will not notice skewness that often signals a gain by transforming to log scales.
- *Don't use categorical variables with numerous nearly empty categories*. Of course, you must notice this problem in the first place.
- *Don't forget to check the conditions for fitting a multiple regression*. Even if the model is complex, you still need to check the similar variance and nearly normal conditions.
- *Don't routinely drop otherwise relevant variables that are not significant*. It is often tempting to drop explanatory variables that are not statistically significant from a regression. Unless the variable is producing substantial collinearity, this step is often useless. Presumably, such a variable is in the model because of substantive issues. While the variable is in the model, you can provide a confidence interval for its slope (even though the interval will include zero). Once the variable is removed, however, all you can say is that it has no significant effect.
- *Don't cheat on the validation*. No fair peeking at the test sample when you are building the model. To be useful, save these data until the end as a test of the model.

BEHIND the MATH

Prediction Errors

The expected value of the sum of the squared residuals in a multiple regression with an intercept and k estimated coefficients is $\sigma^2(n - k - 1)$. This holds if the equation of the model is correct and the observations are independent with equal variance. (Normality is not required.) If that regression model is used to predict independent observations located at the same values of the explanatory variables as the training sample, then the expected sum of squared prediction errors is $\sigma^2(n + k + 1)$. The approximation in the text arises by using two estimates of σ^2. Within the training sample, estimate σ^2 with s_e^2, the square of the standard deviation of the residuals. In the test sample, use the estimator $\sum(y_i - \hat{y}_i)^2/n$.

CHAPTER SUMMARY

Building multiple regression from a large data set relies on substantive insights that suggest explanatory variables for the model. **Segmenting** the data into homogeneous subsets can simplify the modeling task. If one relies on the data to identify the model, as in an automatic search, the model may **overfit** the data, including variables that only coincidentally appear statistically significant. **Model validation** using **training** and **test samples** avoids overfitting. A **partial *F*-statistic** tests the addition of a collection of several explanatory variables to a regression model. The **partial regression plot** shows the data that determine the partial slope of an explanatory variable.

Key Terms

Model validation, 4
Overfitting, 4
Partial *F*-statistic, 18
Partial regression plot, 19
Segmentation, 10
Test sample, 5
Training sample, 5

Objectives

- Know and follow the steps of the modeling process when developing a regression to solve a business problem.
- Convert substantive business insights into explanatory variables in a multiple regression.
- Appreciate the source and impact of overfitting in regression models.
- Use training and test data sets to validate the fit of a multiple regression.
- Know how to use a partial *F*-statistic to test the addition of several variables to a regression model.
- Understand the construction and use of a partial regression plot.
- Recognize the role of automated search algorithms such as stepwise regression for supplementing a substantively motivated model.

Formulas

Partial *F*-statistic

Let R^2_{full} and k_{full} denote the r-squared statistic and number of estimated coefficients in a regression model and let R^2_{partial} and k_{partial} denote the corresponding fit and number of coefficients for a regression that includes a subset of these explanatory variables. The partial *F*-statistic for the excluded subset of variables is

$$F = \frac{(R^2_{\text{full}} - R^2_{\text{partial}})/(k_{\text{full}} - k_{\text{partial}})}{(1 - R^2_{\text{full}})/(n - k_{\text{full}} - 1)}.$$

Partial Regression Plot

Consider a multiple regression of Y on explanatory variables $X_1, X_2, \ldots, X_k$. To obtain the partial regression plot for $X_j, 1 \le j \le k$, regress Y on $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ and save the residuals e_Y from this regression. Then regress X_j *on* $X_1, X_2, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$ and save the residuals e_j from this regression. The partial regression plot is the scatterplot of e_Y on e_j.

EXERCISES

Mix and Match

Match each description on the left to the term on the right.

1. Indicates that a regression model overstates its ability to predict new data	(a) t
2. Procedure for using data to test a model to assess the quality of the fit.	(b) Training sample
3. Scatterplot that shows the marginal slope of Y on X.	(c) k
4. Scatterplot that shows the partial slope of Y on X after adjusting for other explanatory variables.	(d) Partial F-statistic
5. Used to estimate the parameters of a regression model.	(e) Partial regression plot
6. Used to check a regression model for symptoms of overfitting.	(f) Overfitting
7. Number of estimated coefficients in a regression model.	(g) Validation
8. Statistic used to test whether a specific regression coefficient is zero.	(h) Test sample
9. Statistic used to test whether all of the coefficients in a regression model are zero.	(i) Overall F-statistic
10. Statistic used to test whether some of the coefficients in a regression model are zero.	(j) Scatterplot

True/False

Mark each statement True or False. If you believe that a statement is false, briefly explain why you think it is false.

11. Substantive knowledge is very useful when constructing a regression model for a particular application.

12. The baseline category defined by a categorical explanatory variable in a regression should not be a rare category.

13. Overfitting a regression means that the regression has many explanatory variables.

14. The best way to avoid overfitting is to use automatic variable selection procedures such as stepwise regression.

15. A training sample is a random subset of the observations in the original data set.

16. The number of observations in a test sample should match the number of explanatory variables in the studied regression.

17. The partial F-statistic is used to test whether the error variance differs among the levels of a categorical variable.

18. Once you know how much R^2 changes, you can tell whether adding several explanatory variables statistically significantly improves the fit of a regression.

19. A partial residual plot is particularly useful when working with collinear explanatory variables.

20. A partial regression plot reveals whether an outlier influences the partial slope of an explanatory variable.

Think About It

21. An explanatory variable is skewed whereas the response is bell-shaped. Why is it not possible for this variable to explain a substantial portion of the variation in the response in a linear regression?

22. An explanatory variable in a simple regression is bimodal. If the simple regression has a large r^2, then what sort of distribution would you expect for the response?

23. A regression was built to predict the dollar value of a customer transaction. A three-level categorical variable distinguishes cash and credit purchases made in person from online purchases. If this categorical variable explains much variation in the response, what should be apparent in scatterplots of the response versus other variables?

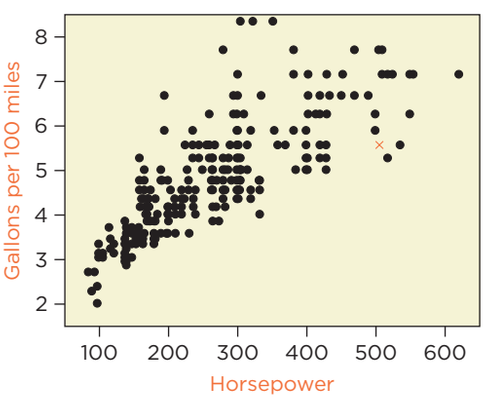
24. Many economic relationships are characterized by diminishing marginal returns. These occur when the effects of increasing inputs (for example, spending for advertising) no longer produce the same magnitude of outputs, such as product sales. If you suspect the effect of an explanatory variable to have this form, what should you do?

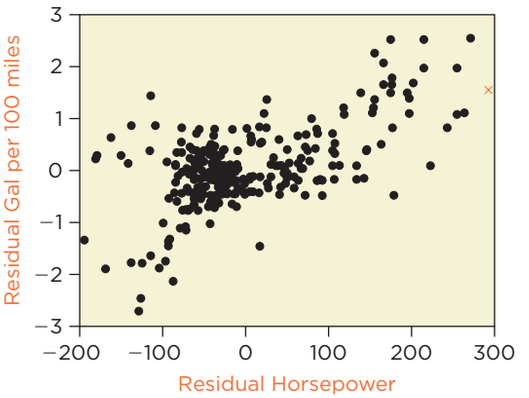
25. A modeler has omitted an explanatory variable that is related to the response but not correlated with the

explanatory variables in the regression. The modeler's regression has $s_e = 135$. The explanatory variable has $s = 15$ and slope $\beta = 3$. How will this omitted variable affect the residuals from the modeler's regression?

26. Why is it useful to identify unbalanced categorical variables before using them in a regression model?

27. An engineer is building a regression that describes the fuel consumption of cars. The following scatterplot graphs fuel consumption (gallons per 100 miles) on horsepower; the partial regression plot on the right is from a multiple regression that includes weight as a second explanatory variable. The highlighted point is a Corvette. Why is the Corvette at the right edge of the partial residual plot rather than the car with the most horsepower?





28. Consider testing the null hypothesis H_0: $\beta_3 = 0$ in a multiple regression with $k = 5$ explanatory variables.
(a) How could you do this with a partial F-statistic?
(b) Because you can also test H_0 with the usual t-statistic, what must the connection be between a t-statistic and partial F-statistic for testing one slope?
(c) Pick a multiple regression. Compare the t-statistic for a coefficient with the partial F-statistic for removing that one variable. How are they related?

29. The estimated slope of an explanatory variable in the training sample is $b_4 = 2.5$ with standard error 0.3. The estimated slope of the same variable in the test sample is $b_4 = 1.6$ with standard error 0.5. Are these estimates statistically significantly different? Explain.

30. Do the partial F-statistic and t-statistics have to agree? For example, consider adding two explanatory variables W and U to a regression. In words, explain why it is possible for the partial F-test to reject H_0: $\beta_W = \beta_U = 0$ but for neither t-statistic to be statistically significant (and hence not reject either H_0: $\beta_W = 0$ or H_0: $\beta_U = 0$). Exercise 41 has a numerical example.

31. In the example of this chapter, the coefficient of the dummy variable for restaurants (identified as Services—Eating and Drinking) is 244.06 when wholesalers define the baseline category in Table 3. When restaurants define the baseline, the coefficient of the dummy variable for wholesalers is –244.06 in Table 4. Why are these estimates the same but for a change in sign?

32. Suppose the explanatory variable X^* is uncorrelated with the other explanatory variables in a multiple regression with response Y. How will the partial regression plot for X^* be different/similar to the scatterplot of Y on X^*.

33. An analyst runs out of time before she can consider the importance of 30 other explanatory variables in a regression. She uses forward stepwise regression to search these, and the first variable it finds has p-value 0.015. Should she use it and look for more? Explain.

34. The variable *State* has 51 levels (including the District of Columbia). If this categorical variable is added to a multiple regression, how many of the estimated coefficients of the dummy variables should you expect to be statistically significant if, in fact, this variable is independent of the response?

You Do It

35. By adding explanatory variables, a modeler doubles the R^2 statistic from 0.25 to 0.50. How does this increase affect the length of approximate 95% prediction intervals? (Hint: Use the approximation $s_e^2 \approx (1 - R^2)s_y^2$.)

36. If the standard deviation of the response in a regression is \$10,000, what value of R^2 is necessary for the approximate 95% prediction interval to have margin of error of \$5,000? (Use the hint from the previous question.)

37. Find the 5% threshold for statistical significance for an F-statistic computed under the following numerator and denominator degrees of freedom.
(a) 1 and 40
(b) 2 and 40
(c) 5 and 40
(d) 5 and 400

38. Find the p-value for the following F-statistics. The two subscripts on each F-statistic denote the numerator and denominator degrees of freedom.

(a) $F_{1,40} = 4.0$
(b) $F_{2,50} = 2.1$
(c) $F_{6,40} = 5.2$
(d) $F_{5,100} = 1.8$

39. The statistical significance of changes to R^2 depends on the size of R^2. In both examples, consider the benefit of adding three explanatory variables to a regression that already contains two explanatory variables. Assume that the regression is fit to $n = 50$ observations.
(a) Is an increase from 10% to 12% statistically significant? Explain.
(b) Is an increase from 90% to 92% statistically significant? Explain.

40. The statistical significance of changes to R^2 depends on the number of variables added to the model. In these examples, the current regression has four explanatory variables with $R^2 = 0.60$ and has been fit to $n = 75$ observations.
(a) Is an increase to $R^2 = 0.61$ by adding one variable statistically significant? Why or why not?
(b) Is an increase to $R^2 = 0.64$ by adding four variables statistically significant? Why or why not?

41. Apple stock Fit the multiple regression of the monthly *Apple Excess Return* on two measures of the overall market, the return on the whole market (*Market Excess Return*) and the S&P 500 index (*SP500 Excess Return*). Use only the 72 months during 2006–2011.
(a) Does the two-variable model explain statistically significant variation in *Apple Excess Return*? Explain.
(b) What is the relationship between the overall F-statistic and the partial F-statistic that tests H_0: $\beta_1 = \beta_2 = 0$?
(c) Is the t-statistic that tests H_0: $\beta_1 = 0$ in the multiple regression statistically significant? That tests H_0: $\beta_2 = 0$?
(d) How can you reconcile the answers to (a) and (c).

42. Apple Stock Compare the partial regression plot for *Market Excess Return* in the multiple regression constructed in the previous question with the scatterplot of *Apple Excess Return* on *Market Excess Return*. Scale the x-axis in the partial regression plot to have the same scale as the scatterplot.

43. R & D expenses These data give ten characteristics of 324 firms. What issues do you see if these ten characteristics are to be used as explanatory variables in a model that describes the salary of the CEO?

44. Financial advising (Chapter data file)
In skimming the marginal distributions (histograms or bar charts) of other explanatory variables in the fee data, what other problems do you find?

45. Financial advising Starting from the model summarized in Table 4, what happens if you add the explanatory variable *State* to the model? (Limit your estimation to the $n = 2{,}000$ cases in the training sample.)
(a) Does the R^2 increase by a statistically significant amount? Briefly interpret your result.
(b) Is any individual t-statistic for a dummy variable representing a state statistically significant at the 0.05 level? How does your answer depend on the choice of the baseline state?
(c) How might you refine the model that includes *State*?

46. Financial advising Dealing with a large collection of categories complicates a regression, and software does not always provide the needed tools. The following procedure is often helpful. Rather than add *State* to the regression (as in the previous question), save the residuals from the model in Table 4. Then use a one-way analysis of variance (ANOVA) to discover whether the average residual in any pair of states is statistically significantly different.
(a) Does the F-statistic of the one-way ANOVA detect a statistically significant difference among the averages? Does the result differ from the partial F obtained by adding *State* directly to the regression? Explain.
(b) Do any pairs of states have statistically significantly different means? With so many comparisons, be sure to use an appropriate adjustment for multiple comparisons (refer to Chapter 26).
(c) How could you incorporate these differences into the regression model without adding the 50 coefficients needed to represent *State*?

47. Financial advising It often becomes tempting to add a large collection of variables to a regression, hoping to let regression "sort out" which variables are important rather than think about why each might matter. Beginning with the model shown in Table 4, add the ten variables that measure population in the ZIP Code of the business location to the model (*Population of ZIP Code* through *Female Population*).
(a) What happens when these are added to the model?
(b) How should these variables be added more sensibly? Does doing so produce a statistically significant improvement in the fit? Explain.

48. Financial advising
(a) Starting with the model in Table 4, add both *Employment* and *Annual Payroll* to the model. Does adding this pair improve the fit of the model by a statistically significant amount?
(b) Instead of adding the pair as in (a), add the ratio *Annual Payroll/Employment* (average payroll salary). Does adding this ratio produce a statistically significant improvement?
(c) Explain your preference among the original model and these two models that add variables.

49. Financial advising The tips from the financial advisers suggest that it may be useful to incorporate the number of credit lines on a log scale in this model rather than a linear scale as in Table 4. Do you agree? Why or why not?

50. Financial advising The following four comments expand the list given in the text on page 9 that suggest six explanatory variables that can be added to the regression model. Use these to improve the fit of the model summarized in Table 4. Be sure to determine whether the effects are statistically significant and consistent with the substantive advice.

(a) The financial provider charges higher fees to risky businesses that have a history of not paying off debt (loans that have been charged off). These clients may not pay the fees charged by the provider, so the provider charges more to compensate for its increased risk.

(b) The longer a business has been around, generally the more skilled it is at managing its money and the less help it needs. The older the business, the smaller the fees.

(c) The financial provider helps manage property leases and charges fees for this service when requested.

(d) Businesses that operate in wealthy locations generate more fees than do those operating in low-income locations.

51. What other suggestions do you have for variables and additional improvements to the model fit in the prior question?

52. 4M

Build a regression for the 891 cases that have had a bankruptcy in the past. The modeling of nonbankrupt clients suggests an initial regression model, but you should suspect that other explanatory variables may be relevant and that coefficients may be different.

(a) *Motivation.* Explain the purpose of the segmentation approach that models these businesses with a history of bankruptcy from the other businesses.

(b) *Method.* Which variables that were not available for modeling a business without a bankruptcy become relevant now?

(c) *Mechanics.* Following the approach used for businesses without a bankruptcy, develop a regression model for predicting fees for those businesses with a bankruptcy. Explain why you did or did not use a test/training sample to validate the modeling.

(d) *Message.* Compare your model with a comparable model for the nonbankrupt businesses. Use examples to illustrate the differences.