SHA573: Using Predictive Data Analysis

What you'll do

Determine the degree of uncertainty in your decision and determine the impact of this uncertainty.
Identify data relationships that can be exploited to reduce uncertainty.
Create a regression model that establishes the impact of two or more attributes on the variable driving the decision.
Refine your regression model to improve its validity.
Create a convincing argument for the validity of your model.
Make a prediction or an estimate using your model.

Course Description

Intelligent business decision making demands you bring all relevant, available resources to bear on a question. Today, the variety of sources and types of data that can aid in decision making is almost overwhelming. The keys to making good use of all these data lie in correctly identifying what to pay attention to and in understanding the relationships between the factors or variables you're monitoring. In order to know, or make good guesses about, what to pay attention to, you need to have experience with the functional details of your organization. When it comes to gaining a deep understanding of the relationships between variables, though, familiarity with statistical methods will provide you with a significant advantage over relying on gut instinct alone.

In this course, you will work through the process of identifying uncertainty in a business decision, choosing variables that might help you reduce that uncertainty, and applying statistical methods to understand both the

relevance of those variables to your decision and the relationships between the variables. By the end of this course, you will have a robust decision model that you can use to make predictions related to your decision. Along the way, you will have clarified and enhanced your understanding of the factors that influence possible outcomes from the decision.

Chris Anderson
Professor, School of Hotel Administration, SC Johnson College of Business, Cornell University

Chris Anderson is a professor at the Cornell School of Hotel Administration. Prior to his appointment in 2006, he was on the faculty at the Ivey Business School in London, Ontario, Canada. His main research focus is on revenue management and service pricing. He actively works with industry, across numerous industry types, in the application and development of revenue management, having worked with a variety of hotels, airlines, rental car and tour companies, as well as numerous consumer packaged goods and financial services firms. Anderson's research has been funded by numerous governmental agencies and industry partners, and he serves on the editorial board of the *Journal of Revenue and Pricing Management* ;and is the regional editor for the *International Journal of Revenue Management*. At the School of Hotel Administration, he teaches courses in revenue management and service operations management.

Author Welcome

As decision makers, almost every decision we make has some level of uncertainty impacting that decision. The goal of this course is to help us reduce that uncertainty. We are going to take our data and model some relationships We're going to quantify those relationships through the use of regression. And then we're going to validate those relationships to make sure that the decisions we've made are in fact better than we would've done otherwise. So let's get into our data and make some decisions.

Table of Contents

Module 1: Discovering Relationships

Module 2: Quantifying Impact

Module 3: Assessing and Validating Your Model

1. Module Introduction: Assessing and Validating Your Model
2. Watch: Accounting for Variable Interactions
3. Watch: Addressing Multicollinearity by Reviewing a Correlation Table
4. Activity: Assess Multicollinearity Using a Correlation Matrix
5. Watch: Using Residual Plots to Detect Nonlinearities
6. Watch: Addressing Nonlinearities Using Transforms
7. Watch: Modeling Interactions between Independent Variables
8. Watch: Considering Missing Variable Bias
9. Course Project, Part Three—Generate a Revised Regression Equation
10. Module Wrap-up: Assessing and Validating Your Model

Module 4: Applying the Predictive Analytics Framework

1. Module Introduction: Applying the Predictive Analytics Framework
2. Watch: Using Your Regression Model
3. Tool: Predictive Analytics Framework Diagram
4. Watch: Testing Your Model with a Holdout Sample
5. Watch: Using Logistic Regression to Model Categorical Variables
6. Watch: Segmenting by Creating Artificial Categories
7. Course Project, Part Four—Validate Your Model
8. Module Wrapup: Applying the Predictive Analytics Framework
9. Read: Thank You and Farewell

1.
2. Excel Step-by-Step Instructions

Module 1: Discovering Relationships

Back to Table of Contents

Module Introduction: Discovering Relationships

When we make projections or decisions based on data, we are doing so with the intent of reducing uncertainty. The value added by data analysis depends on the data you use and your understanding of how best to use it.

In this module, you will begin by developing a mathematical model to represent several factors that influence an outcome or decision. Your initial attempt at modeling most likely will not give you the best achievable model, so you'll refine that model by looking at the importance of each factor. You'll also need to consider the nature of the relationship between each factor and the result in which you're interested. By the time you've completed this module, you will have identified your decision and the factors you believe will influence that decision. This is the blueprint for your predictive model.

Back to Table of Contents

Watch: Using Attributes to Create a Useful Model

As someone with experience in your line of work, you should have a good idea of the factors that influence decisions and outcomes based on decisions that are made in your business context. In fact, you may already collect data on some of these important factors. What may be less clear is how important each of these factors, or attributes, is relative to other attributes. Predictive analysis begins with developing a mathematical model that accounts for all relevant factors.

In this video, Professor Anderson describes the process of building a model, using a fictitious manufacturing business as an example. He describes several important attributes for the example and he introduces the correlation coefficient as a mathematical measure of the significance of attributes.

Transcript

So almost every decision we make is somehow impacted by uncertainty. Our goal is to figure out how to reduce that uncertainty so we can make better decisions. We're going to reduce that uncertainty through discovering relationships. So, you're the sales manager for White Manufacturing. And you're looking to respond to a potential client who has asked you to quote on a prospective job. That job has a series of characteristics and you need to decide, you know, what price to quote them, but at some level when you think that job might be completed and when would you be able to deliver the end product.

So, we go to our operations manager and look for some insight into how long it takes to complete individual jobs. So our operations manager pulls out a sample of recent jobs and looks at the completion time of those jobs. And we notice there's a fairly large range between the shortest and longest completion time. We can calculate the average completion time, you know, it's going to be somewhere in the middle of that range. And we could actually calculate a metric to help us understand how much uncertainty there is in that completion time, we think about that as our standard deviation, which is really the average, about the average, of

those completion times.

So on some level, that average might be the best guest for project completion times but given the size of the standard deviation and the range of potential outcomes, there's a lot of uncertainty in how long it may take to actually complete this specific job. And so, we have this sample of historic jobs, and one of the things we notice from this sample is that each of these jobs has a series of attributes or characteristics. Right? Such things as the number of parts that we were working on, as well as the number of steps or processes through the shop floor that each one of these parts went through. And then to some level whether or not we expedited or rushed this through the shop floor.

Now unfortunately, we don't have an observation, which is exactly the same as this job we're looking to quote. It doesn't have the exact same characteristics as this job we're looking to quote. And so, now we're trying to decide if we can use the characteristics of that sample to help us make a better guess for the completion time for the specific job of interest.

Back to Table of Contents

Watch: Interpreting Correlation Coefficients

Your predictive model is set up to help you focus on one result. In the model, this result appears on one side of an equation. One the other side of the equation is an expression that accounts for all attributes that might influence your decision or outcome. The strength or importance of an attribute can be inferred from a calculated value called a correlation coefficient.

Correlation coefficients are a measure by which you can tell how strongly associated an attribute is with the outcome or decision you seek to understand. In this video, Professor Anderson explains how to interpret correctly the correlation coefficients you will calculate for your model.

Transcript

So, our goal is to try and use these relationships to help reduce the uncertainty in our pending decision. Right? So, one of our questions often is, is well, how do we discover these relationships? So, let's focus on our White Manufacturing example. So, we have a sample of recently completed jobs, and we have this attribute for these jobs which is whether or not these jobs were rushed, or expedited, through the shop floor versus just managed normally.

And, so we can look at the overall average completion time across all our jobs. And then we could also look at these conditional averages, right, or the averages of the subset of sample that was rushed versus the average of the other jobs that were not rushed. And we see a distinct difference in these averages. Right? So here we've discovered our first relationship, that there's a relationship between completion time and whether or not we expedite this job through the shop floor. But the question becomes is, well how do we generalize this away from just simply two categories to a whole spectrum of potential values for that attribute? Right?

So how do we look at relationships between, in essence, two continuous variables, right, where we have our variable of concern, and this other continuous variable, which we think there is a relationship with our

variable of concern. The easiest way to sort of discover those relationships is visually through what we call an XY scatter plot, right? So we're simply generating a graph of our variable of concern on our vertical axis versus our variable which we think is impacting that variable of concern on our horizontal axis. Right? So we think of this as our Y versus our X axis. Right? We talk about the Y axis being our dependent variable. And its value depends on this attribute of interest. Right? So this scatter plot provides a nice visual. We can also summarize this relationship with a simple statistic, which we call our correlation coefficient.

Right? So it provides a sort of single metric that helps us ascertain the strength of this linear association. Right? So a correlation coefficient is a measure of this linear association between our variable of interest and this attribute. If we have a positive association between these two variables, that basically means as one of the variables increases, the other variables also increasing. Right? If we have no, a negative relationship, right, or a negative association, as X increases, Y is decreasing. And then, the other end of the spectrum is really we have no relationship or no association. And so, knowing whether or not X increases or decreases tells us really nothing about what Y is doing. Right?

So we have no association or no relationship. So, this correlation coefficient, this summary measure, is really a measure of consistency and direction, not a measure of steepness. Right? So it tells us how consistent this relationship is. It doesn't tell us how fast or how steep that relationship is. And so, as a measure of this consistency, it has a measure which, or has a value, which ranges from -1 for a strong, negative association, to +1 for a strong, positive association.

And so, if we have no association, then Y is just sort of scattered all over the place for different values of X. If we have weak association, then as X increases, Y tends to increase, but not always. Whereas if we have this strong association then Y always increases as X increases, right? So this correlation coefficient is a measure of strength and direction for this linear association between these two quantitative variables.

Back to Table of Contents

Watch: Recognizing Nonlinear Associations

When you're trying to understand the relationship between a dependent—or decision—variable and your independent—or attribute—variables, one challenge you'll face is that the effect of an attribute may increase or decrease over time. What starts as a negligible effect may come to dominate over time. Or a repeated action may quickly become less effective after the first or second repetition. In situations like these, you can still include the attribute in your model, but you'll need to account for the change. These kinds of associations are called nonlinear because graphs of their effects show curves rather than straight lines.

In this video, Professor Anderson prepares you to recognize nonlinear associations. He also explains how you can transform nonlinear results so they appear as a more usable, linear graph.

Transcript

So when we look at XY scatter plots, or correlation coefficients, to try and unravel our linear associations, that implicitly assumes that for each fixed change in X, or each unit change in X, I'm going to get some fixed change in Y. Right? So regardless if I'm increasing X from 20 to 21, I'm going to get the same impact in Y if I increase X from 120 to 121. That doesn't always working practice. Right? Think about advertising. I spend $100 on advertising, I get a lift in sales. I spend another $100 on advertising, sales increase but not to the same degree as they increased last time. And I spend a third $100 in advertising, again an increase but not as much as the second $100, and definitely not as much as the first $100.

So yes there's a relationship between advertising and demand, it's just not linear. It's some sort of nonlinear relationship. We're getting these decreasing incremental impacts versus fixed impacts over all levels of advertising. So the question is, is how do I uncover these nonlinear relationships? Because if I ignore them, then, and try and sort of look at them as linear, I may underestimate them, or think they don't exist. Right? So, we need to uncover these nonlinear relationships so we can

use those as well in reducing our uncertainty. So, think about demand as a function of time in the market

. Right? So, we're going to have some increased word of mouth the longer our product is out in the marketplace, such that the speed of, or the increase in sales, is increasing in a nonlinear fashion. It's actually increasing as a function of time squared. Right? So demand is increasing rather, rather quickly the longer we're out there. So if we just sort of look at demand as a function of time and look at a linear relationship, we may underestimate the impact of time upon demand. Whereas if we looked at a scatter plot of demand as function of time squared, then we get a very strong linear association, because we've implicitly build in that nonlinear relationship. All right? So we've described that relationship well, given we knew ahead of time, it was nonlinear and it was a function of time squared. But what if we don't know that? What if we don't know it's a function of time squared? How do we uncover that relationship? Well, we can do that using transforms. The most common transform we use is called the natural logarithm.

So, instead of doing a scatter plot of demand versus time, we do a scatter plot of the natural logarithm of demand versus the natural logarithm of time, and it turns out that relationship is linear and strong. And the actual nature of that relationship links right back to the underlying nonlinear relationship, in this case being a function of time squared. We still would have had a line, the line would have been different, different steepedness, if it was function of time to the power of three, or time to the power of four. So let's think about these things in term of demand and price. Right? So obviously, at different points along the price-demand curve the impact of price is going to differ. Right? At lower price points, price is going to drive demand differently than at higher price points.

So we have a nascent, or underlying, nonlinear relationship between price and demand. So we can sort of map out that relationship using our transforms. So now we look at the natural logarithm demand versus the natural logarithm of price. And now we have a linear relationship that we can describe. The nice part about that description is that we get these constant, relative differences versus our constant, absolute differences in our sort of linear world. Right?

In our linear world, we assume that when X went up one unit, regardless of what that was from X from 20 to 21 or 121 to 122, we were going to get the same change in Y. That same thing happens in our nonlinear world, but instead of being a absolute increase, it's going to be a relative or percent increase. So a percent increase in X is always going to generate the same percent change in Y regardless of where we are in our price-demand curve.

Back to Table of Contents

Read: When and How to Do a Linear Transform

- A nonlinear relationship between the dependent (Y) variable and an independent (X) variable occurs when a change in X is not accompanied by a proportionate change in Y.
- Nonlinear relationships are quite common.
- Taking the natural logarithm of data values prepares the data for a linear regression when there is a nonlinear relationship.

It is not enough to show that there is an association between what you are trying to predict (the dependent variable) and something you are able to measure directly (the independent variable). You also need to verify that this relationship is linear before you can draw meaningful conclusions. Linear regression assumes that a dependent variable Y changes the same for each unit change in an independent variable X. This simple assumption may be violated if a change in the value of X causes a proportionate change in the value Y. Nonlinear relationships are actually quite common. Examples include situations characterized by diminishing marginal returns or escalating operational efficiencies. By ignoring potential nonlinearities, you risk overestimating or underestimating an effect over time. The following somewhat-technical deep dive explores the mathematical process of transformation that will help you make use of nonlinear data. Whether or not you fully understand this discussion, you will be able to transform data using tools supplied in this course.

Since linear regressions are relatively simple and straightforward, you will want to use this method when possible. By transforming nonlinear data

before regression, we can prepare it for linear regression without sacrificing validity. For example, take a general nonlinear relationship

$Y = B_0 X^{B_1}$

In this equation, any value of $B_1 \neq 1$ makes the relationship nonlinear. If we take the natural log of both sides of the equation $Y=B_0X^{B_1}$, we get the expression

$$Ln(Y) = Ln(B_0X^{B_1}) = Ln(B_0) + Ln(X^{B_1}) \quad \Rightarrow \quad Ln(Y) = Ln(B_0) + B_1 * Ln(X)$$

Now this general nonlinear relationship between Y and X has been transformed into a linear one between Ln(Y) and Ln(X). If we run a regression on Ln(Y) as a function of Ln(X), we will get good estimates of the parameters of the original nonlinear relationship.

One of the intuitive outcomes that arises from taking the logs of both Y and X prior to running the regression is that the slope of the relationship becomes a measure of elasticity. As a result, this coefficient B_1 represents the percent change in Y for a percent change in X. Recall that in the non-transformed framework, $Y = B_0 + B_1X$ means that if X increased by 10, then Y would increase by $10 * B_1$. If we have a regression now with $Ln(Y) = Ln(B_0) + Ln(X^{B_1})$, it means that as X changes by 10%, we will see a change of $B_1 * 10\%$ in the dependent variable.

When you are accounting for several independent variables, it is possible that only some of them have nonlinear relationships with the dependent variable. In this case, you may only transform some of the independent (X) variables. For example, suppose

$Ln(sales) = B_0 + B_1 * Ln(\textbf{price}) + B_2 * (\textbf{brand})$

where **brand** = 1 indicates a branded product and **brand** = 0 is a private label. In this example, if price changes 1%, then sales will change $B_1 * 1\%$, whereas branded products (**brand** = 1) will have $B_2\%$ higher sales than private label products.

In this last example, the transform for sales and price is referred to as a log-log transform since the natural logarithm was applied to both the dependent and independent variables. The transform for sales and brand is described as lin-log because the logarithm is applied to just one variable while the other remains linear.

Back to Table of Contents

Tool: Log and Semilog Transforms

Semilog and Log-log Transforms

A scatterplot of your data may or may not show a strong association between an independent and the dependent variable in which you're interested. It may be the case that a strong association exists but is masked by the fact that the relationship between the variables is not linear.

Use the Semilog and Log-log Transforms tool, an Excel workbook that allows you to quickly and easily test a single independent variable for association with your dependent variable. To use the tool, open the tool and paste your independent and dependent data into the Log Transforms sheet. The tool will generate a log-log transform in that sheet. At the same time, in the other tab, it will produce data and a visualization for a lin-log, or semilog, transform of the data. You can then inspect these two transforms to see if there is evidence of a relationship between the independent and dependent data. If the data values in one transform or the other appear to form a line, you can infer that a strong but nonlinear relationship exists between the two variables.

Keep a copy of this tool on hand to transform any nonlinear data in two columns, where the first column is an independent variable and the second column is an associated dependent variable.

Back to Table of Contents

Case Study: A Revinate Study Shows Nonlinear Impacts of Engaging with Consumers

- **If the impact of an independent variable is nonlinear, a regression might indicate an erroneously weak or nonexistent association with the dependent variable.**
- **Where a nonlinear effect is suspected or hypothesized, it is sometimes helpful to adjust (transform) the data before performing the regression.**

TripAdvisor is a business that allows consumers to post hotel ratings and reviews online and also gives the hotel owners an opportunity to post responses to individual consumer reviews. Here is an example of a comment made by a guest, followed by a response from the hotel.

A hotel might ask what benefit, if any, is there in engaging with guests' online posts. Will these responses have a significant impact on future bookings and guest attitudes? Since most hotels will receive numerous reviews daily, the time and effort to respond to guest feedback in a personalized way can have a significant cost.

One might hypothesize that by responding to some reviews and indicating what efforts the hotels had taken to address guests' issues, hotels would create a perceived benefit to potential guests. This positive perception might translate to an increase of future bookings. Likewise, it may be thought that simply thanking consumers for their review may not add value to future guests. To investigate these potential effects, we looked at the revenue generated by hotels from consumers booking

rooms at TripAdvisor as a function of the percentage of reviews with a response from the hotel. We did this in conjunction with online reputation management firm Revinate (see revinate.com for company background) using the Revinate Surveys tool. Revinate Surveys is a survey product that allows hotels to collect private and public guest feedback simultaneously. The survey product allows hotels to send a post-stay short-format survey to guests.

This figure illustrates the relationship between revenue and rate of hotel responses as a percentage of guest reviews on TripAdvisor.

The figure indicates that responding to some reviews increases revenue, and that revenue is maximized when hotels responded to about 40% of guest reviews. But the figure also indicates that responding to all reviews is actually worse than responding to none. The impact of hotel response (X variable) upon revenue (Y variable) is positive and then negative—that is, it changes as the value of X changes. So if we had simply used regression without adjusting for our hypothesized nonlinearity, we would have most likely found no relationship between responses and revenues. For the full details of how we performed the regressions and details of the analysis, see the full report.

To read the full published report of the Revinate study, download the Hotel Performance Impact of Socially Engaging with Consumers PDF.

Back to Table of Contents

Identify Variable Relationships

In this quiz you will practice identifying the nature of relationships between independent and dependent variables.

Scenario

As a commercial real estate developer, you are looking to get some insight into what affects the selling prices of commercial properties. In an effort to estimate the value of a pending property, you have put together a sample of recently completed transactions. For each of these transactions, you have the following properties:

Selling price
Number of parking spots
Size of building in square feet
How many months ago the sale was completed
Number of cars per hour that drive by the location
Desirability of location, on a scale from 1 to 10, where 1 is most desirable

If you were going to build a model, what would be the dependent variable?

Of the remaining variables, indicate a variable you think might be linearly related to the dependent variable and indicate why.

Back to Table of Contents

Analyzing in Your Area of Expertise

Discussion topic:

Effective analysis relies on familiarity with factors that may influence outcomes in your business sector. For instance, if you want to predict the selling prices of commercial properties, it helps to have spent some time buying and selling commercial real estate, or at least having worked in some capacity that allowed you to develop some hypotheses about how commercial real estate works.

How familiar are you with the important details of how your organization operates? By reflecting on and discussing your strengths and weaknesses, you will be better equipped to embark on predictive analysis. You'll know where to solicit input, and by interacting with your fellow students in this class, you may gain insight into your situation.

Create a post in the discussion board in which you:

Summarize the experience you bring to decision making or prediction in your area of expertise, focusing on specifics that make you particularly well-suited to identifying factors that influence how things work in your organization or business sector.
Identify areas where you feel you could improve or expand your familiarity with your business sector, or where you feel you could benefit from soliciting input or feedback from others.
Describe at least one action, apart from completing this course, that you could take to improve your readiness to apply predictive analysis effectively in your area of expertise.

To participate in this discussion:

Use the **Reply** button to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of

someone else's work as your own without source credit).

Back to Table of Contents

Course Project, Part One—Consider How a Variable Impacts Your Decision

In this part of the course project, you will consider a question or decision and begin to apply a predictive analysis approach to that decision. In this first part of the project, you will focus on identifying independent variables that may influence outcomes. *Completion of all parts of this project is a course requirement.*

Instructions:

1. Download the course project document.
2. Complete Part One.
3. Save your work.
4. You will submit your completed project at the end of the course for grading and credit.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrap-up: Discovering Relationships

Predictive analysis is built on a foundation of good data and a working familiarity with the situation for which data is being collected. As a first step, you have identified a business situation or decision you are interested in modeling and have begun the process of identifying the factors, or independent variables, you believe may shed light on that decision.

In the process of selecting which independent variables to consider, you began to think about their possible relationships with the variable of interest. Are these strong associations, or weak? Is the change in your variable of interest linear with respect to a given independent variable, or not? As you continue through the course, you will examine these relationships more closely and continue to develop a regression model that you can refine and then put to use.

Back to Table of Contents

Module 2: Quantifying Impact

Back to Table of Contents

Module Introduction: Quantifying Impact

For any given decision or prediction you make, the number of factors that could potentially impact the outcome is nearly infinite. Experience tells us that very few of these factors have any significant impact in practice. It's impractical to account for any but the most significant factors.

Now that you have identified factors you believe will influence your decision or prediction, it's time to develop an initial model that allows you to begin understanding in statistical terms how each factor is related to your outcome. The choice to include a data stream in your model amounts to a hypothesis that this data describes a factor that is significant for your decision. Part of the process of building your model involves testing each of these hypotheses, one by one. You will do this by running regressions, from which you will be able characterize the relationships of each independent variable with the dependent, or decision, variable. You will also test mathematically how the independent variables might interact with one another.

Back to Table of Contents

Watch: Describing Relationships with a Scatterplot and Best Fit Line

You can think about the strength of the relationship between two variables in terms of a correlation coefficient, and a scatterplot helps you express that relationship visually. By looking at a scatterplot, you can easily see whether an increase in X corresponds to an increase or decrease in Y, and you can also see about how much Y changes for a given change in X. As a next step, you'll want to create a mathematical equation that captures this relationship. This equation defines a best fit line for the data in the scatterplot. The best fit line follows the trend described by the data points on the scatterplot. It helps you predict, for a given value of X, what value you can expect for Y.

In this video, Professor Anderson explains how and why a best fit line is useful. In particular, he introduces the notion that employing a best fit line can reduce the amount of uncertainty inherent in any decision or prediction.

Transcript

So we can use correlation coefficients and scatter plots to help us understand the existence of potential linear relationships between variables and ultimately assess whether or not we have any underlining nonlinearities. Our next step is to actually quantify the structure of that relationship and physically measure how fast Y changes given changes in X. We're going to do that through the use of regression.

So we have some variable of concern, Y. We have an average for Y, and that's a pretty good guess at future values of Y. Fortunately for us, we have an attribute X, and we can use that attribute X in the relationship between X and Y, to sort of refine our guess for Y. The question is, is what is the correct structure of that relationship? What is the equation of that line that maps out that relationship? So when we overlay our line on our relationship, we'll notice that the data points don't necessarily fall on the line, but there's a difference between the data point and the line for each corresponding X.

We want to find the line such that that difference between the data point and the line is minimized, and specifically we're going to take all those differences and square them, sum them up, and then minimize the sum of those squared differences. That will be our line of best fit. One of the interesting things to notice, is once we have this line of best fit, how much better that line is at predicting values of Y, than simply using the average for Y. In an essence, each point along the line is a conditional average for Y given X. And we can quantify how much better that conditional average is than using our original average.

So, if we look at the deviation, or the difference, from the original average to each data point, that's our underlying measure of uncertainty in using the average as a future prediction. Whereas, if we look at the difference between the data point and the line, we also have this difference, we'll notice that those difference are smaller. And the degree to which that difference is smaller, we've explained part of that difference away.

And in regression, we have this thing called R-squared, where we're basically talking about how much of that deviation we've explained the way by using R line. And so R-squared is a percentage of the deviation that's explained. So we're going to fit this law into our data points, we're going to fit that line to minimize the sum of those squared errors. Once we've done so, we can talk about how much better that line is at predicting Y as a function of X, than just using Y alone through this R-squared as a measure of the percent deviation explained.

Back to Table of Contents

Watch: Obtaining the Equation for a Best Fit Line

In practice, there are many ways to derive an equation for a best fit line. In Microsoft Excel, with all your data available, the process of getting this equation is very straightforward.

In this video, Professor Anderson works through an example in which he shows two ways to arrive at a best fit line equation in Excel. In this example from the White Manufacturing scenario, the equation of the best fit line describes the relationship between number of pieces in a job and completion time. This estimate based on a linear relationship is already a great improvement over estimating completion time based solely on the average time to complete all jobs.

Transcript

So, let's look at an example of how to find a best fit line in Excel. So, White Manufacturing is try to ascertain the impact of the number of pieces in a job upon the completion time. So we could generate a correlation coefficient for completion time and the number of pieces. We could look at a scatter plot between completion time and the number of pieces to ascertain whether or not we have a relationship, and to what degree that relationship is linear as we investigate that with our scatter plot.

Now we want to quantify that relationship to get a sense of the magnitude of the impacts, given our assessment that there is a relationship. So there's a couple of different ways we could do this in Excel. We could use the slope intercept and R-squared functions to actually generate the structure or characteristics of the line and to what degree it fits the data.

We can also generate those characteristics of the line right in our scatter plot by asking Excel to display the equation of the line and the R-squared in our scatter plot. Once we've done either of those approaches we get an equation for completion time, which basically indicates that the completion time is some fixed component, plus some function of the number of pieces in that job. And so now we have a new estimate for the

average completion time, which will be a function of the characteristics, in this case the number of pieces in that job.

Back to Table of Contents

Tool: Least Squares / Best Fit Model of Regression

Least Squares / Best Fit tool

An intuitive understanding of the principles of the least squares model of regression will help you appreciate the significance of the results of your regression analysis.

Use this tool to estimate the slope and intercept of a best fit line. Follow this sequence to familiarize yourself with the tool.

1. Begin in the **data** tab. Enter four pairs of values for the **time** and **data** variables.
2. In the **line** tab, adjust the values for intercept and slope until you feel you have an approximate best fit line for the points shown.
3. Now look at the **errors** tab. Notice the error bars between each line point and the estimated best fit line.
4. Finally, use the sliders in the **squares** tab to fine-tune your estimate by minimizing the total area of the four boxes formed by the vertical error bars.

You will use this tool in an activity that helps you develop intuitive familiarity with how the least squares concept works. After completing step 2, the line you see will be a decent approximation, but it will probably not be the best fit line. As you adjust the sliders in step 4 you can methodically close in on the parameters for the best fit line. Remember, in daily practice Excel will calculate this line for you. This purpose of this activity is just to give you a feel for what best fit means.

Back to Table of Contents

Activity: Estimate Slope and Intercept for a Best Fit Line

Least Squares / Best Fit tool

In this activity, you will practice estimating the slope and intercept of a line by visually inspecting and adjusting a line of best fit. Completing several repetitions of this exercise will help you develop an intuitive feel for the meaning and significance of best fit lines. Open the Least Squares / Best Fit tool and follow these steps to complete the exercise for each of the examples in this activity.

1. Enter the points into the **data** tab. For each pair of numbers, enter the first number in the **time column and enter the second number just to its right in the data column.**
2. In the **squares** tab, adjust the sliders until you feel the line most closely approximates the best fit line for the four points shown.
3. Reveal the Example solution on this page, and compare the values for **intercept and slope with the values you estimated in the tool.**
4. If necessary, use the sliders to adjust so the values in the tool match the values in the solution. As you do so, notice how the best fit line changes with respect to the four points.

If you haven't already downloaded the Least Squares / Best Fit tool, you can get it using the link above.

Example 1 points: (1, 1); (2, 2.5); (3.5, 3), (5, 4)

Example 1 solution

intercept = 0.65; slope = 0.69

Example 2 points: (1, 2.5); (1.5, 3.3); (3, 5.2), (3.5, 4.8)

Example 2 solution

intercept = 1.67; slope = 1.01

Back to Table of Contents

Watch: Testing for Statistical Significance

It is not enough to establish a correlation between an independent variable and your dependent variable. You need to be able to say with confidence that the association you have identified is meaningful. Statistical significance is a critical, mathematical measure of the relevance of a result. For any coefficient that defines the slope of a best fit line, you will also want to calculate a p-value, which is a probability that essentially allows you to express a degree of confidence in your result. For example, you could hear a statement like "Based on our data, we are 95% confident that the 60-day repurchase rate will be between 25% and 30%." Here, the 95% confidence, which is based on a p-value, expresses statistical support for an outcome, but not certainty. So, repurchase rate might be less than 25% or greater than 30%, but the statistical evidence supports a result between these two values.

In this video, Professor Anderson explains how hypothesis testing works and how to think about statistical significance in the context of regression lines.

Transcript

All right, so we've put together a XY scatter plot of our variable of interest, completion time, versus the number of pieces that are a part of our job. And then we fit a regression line to that getting the equation of the line and the resulting R-squared. We could also generate that same regression model in other methods within our spreadsheet or other statistical packages. So here we have output from Excel generating those regression parameters in a slightly different format.

Here you'll see we have our R-squared. We have our intercept and we have our coefficient for pieces. But we also have a lot of other information. A key piece of information that we're going to focus on here are the P-values for both the intercept and the pieces coefficient. These P-values, or probabilities, are statistical measurements of the significance of these coefficients. We have to realize that we have this sample that we built our regression model with and this is just a small subset of all

possible jobs that we might look at building regression models on. And we're trying to use this sample to make inference about a future job.

And so, this model, we're trying to generalize from this model about all past, present, and future circumstances for our shop floor. And so, because of that, we have to be cautious about how definitive of statements we make based upon the parameter estimates that come from this sample. So yes, they are a fact based upon the sample, but we're trying to generalize from the sample to the overall population of potential values. And so, in our statistics world, we typically have this thing called hypothesis testing, where we kind of setup the opposite result we're trying to find and we basically prove that the opposite result is not true, which provides support for the thing we're looking for.

So in this circumstance, we're trying to find a relationship. So the opposite is that is that there's no relationship. So that would mean that the coefficient for pieces as part of our regression equation is 0. And what we simply do, is, statistically we estimate the probability that given the sample we have and the estimate from the sample, that the population value could be zero. And here, for our sample, that probability is really, really small. And so, basically from that we can gauge that there is no chance or very little chance that there is no relationship, which provides support for the relationship that we found. Right? And sort of adds confidence to us that this model is valid and in our statistical world, we say that we have statistically-significant results. Right?

So we have some model, that is, we can reject the idea that there is no model which provides support that we have some model. Right? And so, one of the things we look at from either regression based output from Excel, or any statistical package, is we look for this thing called P-value, or probability. And we basically want that number to be less than 0.05. Right? So we have this 5% critical level. So as long as there's less than 5% chance, that there is no relationship, then that means we are comfortable saying there is a relationship. Right? We are not saying that it's exactly this but we are saying, in our case our parameter estimate is not zero, which provides support that keeping pieces in our model helps us predict completion time and reduces that uncertainty in the overall completion time.

Back to Table of Contents

Activity: Run a Single Variable Regression on Data

Consider a firm that is looking to determine sales of new products. The firm typically launches products in small test markets prior to a broader launch. The firm wants to use regression to describe a relationship between test market sales and full launch sales. The slope of the best fit line will describe this relationship, and the r-squared value will give an indication of how effective the market test is as a predictor of sales after launch. In this case, a higher r-squared value indicates that test sales have a better predictive value.

In this activity, you will practice running a simple regression for one independent variable compared against a dependent variable of interest.

To complete this activity:

1. **Download the sample data workbook.**
2. **Download and open the Create a Regression How-To Guide.**
3. Open the workbook and follow the instructions in the how-to guide to create a regression using the data in the **data** tab of the workbook. Your result should be expressed as an equation of a best fit line for the data.
4. When you are done, compare your results with the answer tab in this workbook. If you obtained unexpected results and can't figure out why, get in touch with your instructor.

Your practice running a regression on this sample data will prepare you to complete a similar task with your own data in the course project.

Back to Table of Contents

Watch: Using Multiple Regression to Consider Several Relationships Together

A regression report summarizes the association of an independent variable with a dependent variable of interest. This association is expressed as the slope and intercept for a best fit line through a data plot along with an r-squared value. Essentially, this report answers two questions.

As we see a change in the factor we are watching, how is our outcome changing?
How reliable is a change in the outcome, given a change in the factor we are watching?

But we're rarely interested in the impact of just one variable factor on our outcome. While it's not necessary to account for every variable, an oversimplified model will rarely provide meaningful results. What we need is a regression that considers all relevant variables together.

In this video, Professor Anderson works through an example of a regression that accounts for multiple attributes, providing a slope coefficient for each attribute. The calculations that consider two or more variables at the same time is referred to as a multiple regression.

Transcript

As White Manufacturing, we're trying to get an estimate for completion time for this pending job. So we collect a sample of recently completed jobs to get a sense of if we can find some relationships between the attributes of those jobs and completion times. If we can find those relationships, then we can have a more refined estimate for this pending job. For example, we look at whether or not jobs were rushed or expedited through the shop floor. And we look at the average of those jobs that were rushed versus the average of those jobs that were not, and we see a difference in those completion times, providing some insight that this attribute is going to reduce the uncertainty in our estimate of completion times.

We have two other attributes, the number of pieces and the number of steps, which are quantitative variables. And we can look at the correlation between those quantitative variables and our completion time to get a sense if there's any linear associations. We can verify those linear associations through XY scatter plots of completion times versus the attributes. And then we can overlay regression equations to help us quantify the impacts of those linear associations. But here we've looked at those three attributes in isolation. We should go one step further, given there may be some potential interactions between those three attributes, and look at the impact of those three attributes together upon completion time.

So we're going to run a regression now but with multiple attributes. This is what we refer to as multiple regression because of those multiple attributes. So we generate our regression output. This regression output looks the same as our single variable regression. We just have estimates for all three of our attributes versus estimates for a single attribute at a time. We do notice that the R-squared from this multiple regression is considerably higher than the individual R-squareds for our models in isolation, indicating that this composite model is doing a better job at reducing uncertainty than either of those individual models.

Now, when we look at our coefficient estimates, we want to confirm that the P-values are significantly small, less than .05, indicating that statistically speaking there is a relationship and that these coefficients are not zero. The one thing we notice though is that these coefficients in our multiple regression are a little bit different than those corresponding coefficients in our simple regressions or our single-attribute regressions.

For example, if we looked at the coefficient estimate for whether or not a job was rushed that estimate is a little bit different than the difference between the average completion times for jobs that were rushed versus the average completion times for jobs that were not rushed. So when we look at these things together, we get a, sort of more refined estimate of their joint impact. We can take those joint impacts to generate an estimate for completion time as a function of the job characteristics and these coefficient estimates.

The one thing we notice though is that we have this negative value for

our intercept. Right? The intercept is the part of our regression equation which is independent of the product attributes. And so we have to ask ourselves does it make sense to have the component of completion time, which is independent of the product attributes, to be negative, or if it doesn't make sense then perhaps we need to revisit our model.

Back to Table of Contents

Activity: Run a Multiple Variable Regression on Data

Consider a firm that is looking to determine sales of new products. The firm typically launches products in small test markets prior to a broader launch. They want to use regression to help them determine the relationship between test market sales and full launch sales. *In addition, they would like to determine the impact of sales price on actual sales.*

In this activity, you will practice running a multiple regression for two independent variables, **test sales** and **selling price**, compared against final sales, the dependent variable of interest.

To complete this activity:

1. **Download the sample data workbook.**
2. **Download and open the Create a Regression How-To Guide.**
3. Open the workbook and follow the instructions in the how-to guide to create a regression using the data in the **data** tab of the workbook. Your result should be expressed as an equation of a best fit line for the data.
4. When you are done, compare your results with the answer tab in this workbook. If you obtain unexpected results and can't figure out why, get in touch with your instructor.

Your practice running a multiple regression on this sample data will prepare you to complete a similar task with your own data in the course project.

Back to Table of Contents

Read: Coding Categorical Independent Variables

- Coding categorical variables using a sequence of numbers can suggest a progression that is not valid for the data.
- Using binary variables that can be set to 0 or 1 for each category avoids false numerical associations within the categories of a categorical variable.

Regression by its nature is a quantitative tool. But sometimes we are looking at categorical data, such as gender or political party affiliation. When performing regressions involving categorical independent variables, we can code categories into numbers. For example, if our categorical variable is gender, we could code female = 0 and male = 1. If we are trying to determine the impact of gender on salary, we could use the equation

Salary = B_0 + B_1* GENDER

that has a numerical solution. By choosing to code with 1s and 0s, we simplify the interpretation of gender in our equation. For females, coded as 0, the coefficient B_1 disappears and has no impact on salary. This makes it easy to isolate the incremental impact of gender on salary since B_1 represents the additional salary paid to males.

Now consider another scenario in which we're concerned with size categories. These could be coded small = 0, medium = 1, large = 2. So we are solving for dependent variable Y with the equation

Y = B_0 + B_1 * SIZE

If we were to run a regression using 0, 1, and 2 to represent size categories, our result would implicitly dictate the relationship between the three values. The impact of selecting large over small would appear to be twice the impact of selecting medium over small. This is a problem because our variable choice is dictating the relationship instead of allowing the regression to determine the relationship. In fact, any other arbitrary choice of numbers to represent the categories (small = 1, medium = 3, large = 5) would dictate a different relationship between the categories.

We can let regression determine the relationship between the categories by coding again with 1s and 0s. In this case, we'll need to create some additional variables. We can create a variable for Large that equals 1 if our size is large and equals 0 otherwise. A variable for Medium is 1 for a size medium and 0 otherwise. In this case, the two variables are sufficient since when both these variables are 0 (not large and not medium), then the size must be small. So we need *n* – 1 variables to model a categorical variable with *n* categories. Our equation in the size example would look like this:

$Y = B_0 + B_1 * \text{Large} + B_2 * \text{Medium}$

So for size small, $Y = B_0$. For medium, $Y = B_0 + B_2$. And for large, $Y = B_0 + B_1$.

By coding these as 1s and 0s, the regression itself can determine the impact of the different categories (through the Bs) versus us presupposing it via our coding of 1, 2, 3, etc.

Back to Table of Contents

How Available Is Your Data?

Instructions:

You are required to participate meaningfully in all discussions in this course.

Discussion topic:

Predictive analysis requires you to have the key data available to build a model. As you consider your data needs, create a post in which you answer the following questions.

What challenges do you anticipate as you seek to acquire the data you need, and how (if at all) is this expectation informed by prior experience? How can you address the issues that may prevent or impede your access to data?

To participate in this discussion:

Click **Reply** to post a comment or reply to another comment. Please consider that this is a professional forum; courtesy and professional language and tone are expected. Before posting, please review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Course Project, Part Two—Map Decisions to Outcomes

You are now familiar with multiple regressions and have practiced running a multiple regression on data. Now it's time to apply this to your own problem of interest. In this part of the course project, you will define how values of independent variables relate to the variable of interest. You'll build a regression model and consider whether the regression equation accurately captures the relationships between the independent variables and the dependent variable. *Completion of all parts of this project is a course requirement.*

Instructions:

1. Open your saved course project document. (If needed, download it again now.)
2. Complete Part Two.
3. Save your work.
4. You will submit your completed project at the end of the course for grading and credit.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrap-up: Quantifying Impact

You began this module with a question or decision and with some tentative hypotheses about what factors might impact that question or decision. You have progressed to having a fully developed regression model that considers multiple factors and expresses each factor in terms of numerical values and statistical significance. This important step toward making a scientific, data-driven decision is a real accomplishment. Already you are considering your decision in a way that is likely more thoughtful, and is certainly more defensible, than before you began this process.

If you were to stop here, however, there would be a good chance of distortion in your decision-making model. In the next module, you will validate and refine your regression model.

Back to Table of Contents

Module 3: Assessing and Validating Your Model

Back to Table of Contents

Module Introduction: Assessing and Validating Your Model

It's important to remember that the results of a regression analysis are merely numbers. In order to create valuable insight from these numbers, you will need to use your business intuition to interpret them. Often the numbers in the first regression analysis will yield contradictory or unclear results. You will need to decide what to do in this case, and you may choose to modify the regression model.

In this module, you will test your regression model for validity. You'll identify possible interactions between two or more of your independent variables, and you'll use correlation tables to see how to adjust your model to account for these interactions. You will generate residual plots to help you more readily identify nonlinear associations, and you'll adjust for these nonlinearities. When you have completed the actions outlined in this module, your resulting model should be significantly more robust and a much better predictor of outcomes.

Back to Table of Contents

Watch: Accounting for Variable Interactions

At this point you have identified some factors (independent variables) that have an impact on the outcome (dependent variable). As you review your multiple regression, look at the values of the coefficients for each of the independent variables as well as the intercept value. Are there opportunities to improve the accuracy of the model?

In this video, Professor Anderson considers a situation in which the values in the regression equation hint at a need to improve the model. He introduces a new synthetic variable to resolve an issue with the behavior of the model.

Transcript

As part of our regression modeling process, we want to make sure that the resulting model we've produced makes sense. That it's logical, it translates to our physical setting. So a company, White Manufacturing, has put together a regression model based upon a sample of recently completed jobs. It's going to use this regression model to help it predict future completion times, as a function of the characteristics of the jobs that go through its shop floor.

The characteristics of those jobs are the number of pieces that are, they're working on, the number of steps that that job has to go through in the shop floor, and whether or not that job is expedited, or rushed, through the shop floor. So we've built our regression model. We have our coefficient estimates. The P-values for those coefficient estimates are all less than 0.05. We have statistical support for our model. It has a R-squared, which is indicating a significant reduction in uncertainty in completion time.

And so at some level, we're very happy with our model. But our model has this intercept, which is negative. That intercept indicates that the part of completion time which is independent of the job characteristics is negative. And so at some level, physically, that doesn't really make sense as part of how we understand what we work on, right, or how we

understand what goes through our shop floor. And so perhaps we need to revisit that model and refine that model to make sure that our coefficient estimates are logical.

So, having deep understanding of what we do at White Manufacturing, we realize that if a job has 100 pieces and goes through 10 different steps on the shop floor, that means we've worked on that thing 1,000 times, right? 10 times 100. And so, that's very different than if we look at pieces and number of steps in isolation. So, we go back to our regression modeling process, and we create this new variable, which is the product of the number of pieces times the number of steps. Right? So we've created this variable which is a function of these other variables in our sample.

Now we look at a scatter plot of that new variable where we have completion time as a function of the product of pieces times the number of steps. And here we see a very strong positive linear association, when we put that variable back into our regression model. And so, now we look at the impact of the number of pieces, the number of pieces times the number of steps, the number of steps and whether or not this job was expedited or not. We look at all four of those attributes together, we get an even higher R-squared than before. And we have statistically-significant coefficients. Your P-value is less than 0.05.

And so, we have an even better model than we had previously. And now, our coefficient for our intercept is positive. Right? So, before we had this sort of nonsensical coefficient, that coefficient turns out to be positive now. So, we've sort of fixed that prior problem with this stronger model. But we've introduced another issue because now when we look at the coefficient for pieces, it is actually now negative, right? Before it was positive, now it's negative and it doesn't makes sense for a job that has more pieces to take less time. And so, we have potentially introduced another issue in our model. And that issue is something we refer to as multicollinearity.

Back to Table of Contents

Watch: Addressing Multicollinearity by Reviewing a Correlation Table

Sometimes one of your independent variables has a significant impact on the values of another independent variable, particularly if you begin introducing synthetic variables. When two attributes interact, this can distort your model since you are in essence accounting for the influence of one attribute in more than one place in the model.

In this video, Professor Anderson examines a case of interacting independent variables and shows how this situation can be addressed. He uses a correlation table to examine in detail how strongly variables are related.

Transcript

When putting together a multiple regression model, one of the issues we want to be aware of is the potential for multicollinearity. So keep in mind, the purpose of regression modeling is to exploit relationships between the independents, and the dependent variable. Right? We're trying to refine our estimated dependent by using relationships between that dependent and the independents.

Multicollinearity arises if there's strong relationships between the independents themselves. Right? So now we're looking for strong relationships between independents and dependent, but if the independents themselves are strongly related, then our model can get mathematically confused. So let's look at an example. White Manufacturing is trying to reduce the uncertainty in its estimate of completion time. It's going to do so by using three attributes of the jobs. The number of pieces, the number of steps and whether or not that job was expedited, or rushed, through the job floor.

To improve our modeling efforts, we've created this new synthetic variable, which is the product of the number of pieces times the number of steps. Inclusion of this fourth attribute into our regression model has created a very strong regression model that explains a lot of uncertainty

in completion times. It has coefficient estimates with P-values less than 0.05. So, statistically speaking, we have a really good model. Unfortunately, one of the parameter estimates is a little nonsensical, that is the coefficient for pieces is negative, which indicate that the more pieces we have in a specific job the less time it should take. And that's sort of counter intuitive. The reason we get this counter intuitive coefficient is because of multicollinearity.

So, let's look at the correlations between our four variables. Right? The three attributes and the synthetic variable, and our dependent variable, completion time. And, a couple of things arise, we see that the correlation between this synthetic variable and pieces is actually bigger than the correlation between pieces and our dependent variable completion time. Right? So the correlation between the attributes is stronger than the correlation between one of those attributes and the dependent variable. Right? That's going to be an issue. Right?

So because of that, we can no longer include both of those attributes in our regression. We're going to want to include whichever one of those attributes which is most strongly related to our dependent variable. That turns out that the correlation between the synthetic variable is stronger, so that the correlation between that synthetic variable and our dependent variable is stronger than the correlation between pieces and our dependent variable. So, we'll keep that new guy, and throw out the pieces one.

Conversely, when we look the number of steps, we see that the correlation between the number of steps and the synthetic variable, pieces time steps, is actually smaller than the correlations between steps and time, or between pieces and steps and time. Right? So that means even though that this synthetic variable is a product of the number of steps, we can keep both of those attributes in our regression model. Right? So we don't have a potential multicollinearity problem.

So we can rerun our regression now with these three attributes, right, pieces, time, steps, the number of steps, and whether or not this job was expedited or rushed. And we noticed that the R-squared here is only slightly smaller than when we had all of four attributes in our regression model. We still have P-values for these individual coefficients which are

less than 0.05. And, all the parameter estimates make sense. Right? They're all, if they're supposed to be positive they are positive. If we think they should be negative, they are negative. Right? So we now have this model which, statistically speaking, makes sense. And practically speaking, also makes sense.

Back to Table of Contents

Activity: Assess Multicollinearity Using a Correlation Matrix

You can use a correlation matrix to detect possible multicollinearity among independent variables in your regression model. In this activity, you will create a correlation matrix based on data provided. You will use the correlation matrix to compare the strength of associations among independent variables to associations with the dependent variable. If you find associations that are strong enough to distort the results of your model, you will call out one or more independent variables as candidates for elimination from your model. You can download a work aid to help you complete this activity from the link above.

To complete this activity:

1. **Download the sample data workbook.**
2. **Download and open the Create a Correlation Matrix How-to Guide.**
3. Open the workbook and follow the instructions in the how-to guide to create a correlation matrix using the data in the **data** tab of the workbook.
4. Once you have obtained the correlation matrix, use a series of comparisons of correlation coefficients in the matrix to determine if there are independent variables that are candidates for exclusion from the regression model. Use the following sequence to make your determination.
   - Trace rows and columns to identify the value of correlation between any two variables.
   - For any pair of independent variables X_i and X_k, is the correlation between X_i and X_k of a higher value than either X_i's or X_k's correlation with the dependent variable Y? If so, make a note of which X has the higher correlation since this X could be a potential candidate for exclusion.
5. Record the results of your correlation matrix in the workbook. Besides the completed correlation matrix, your answer should include both a conclusion based on your findings and an explanation of why some or no independent variables are candidates for

exclusion.

6. When you are done, compare your results with the answer tab in this workbook. If you obtained unexpected results and can't figure out why, get in touch with your instructor.

Your practice creating a correlation matrix from this sample data and analyzing it will prepare you to complete a similar task with your own data in the course project.

Back to Table of Contents

Watch: Using Residual Plots to Detect Nonlinearities

Your model should contain only attributes that help your prediction. In some cases, an attribute (independent variable) is strongly associated with the dependent variable but that association is not linear. You want to be able to use the predictive power of the attribute, but you will need to account for the nonlinearity. The first step, though, is recognizing that there is a nonlinear relationship. How can you do this?

In this video, Professor Anderson introduces the residual plot, which makes it easier to recognize nonlinear relationships between attributes and the outcome. He shows an example of a residual plot for which it's easy to see that a nonlinear association exists.

Transcript

In regression, we refer to the difference between the estimated value from our model and the value that actually happened as our residual. Right? This residual is the part that's left over from our model, the part that we can't predict. In an ideal world, we want that part that we can't predict just to be some random noise, right, just to be stuff that's not a function of other things.

So, let's look at a simple example. Here we have a two variable setting, where we've hypothesized a relationship between Y and X. We've done a scatter plot, of Y as a function of X, we've overlayed our regression equation on that scatter plot. And so, that regression equation represents our prediction, and we can see the differences between that prediction and a series of our data points. And we see that that difference is changing as the attribute increases. Right?

At one level, the prediction is less than our sample at another range, we flip to be our estimated value greater than the sample and then we go back to being less than the sample for higher values of our attribute. So the residual seems to be displaying some kind of pattern or function of, in this case, the underlying attribute. So we can look at those residuals in a

residual plot and what a residual plot is, is simply just those differences but displayed as a function of the predicted Y. Right? So it makes the visualization of those residuals much cleaner. And so, now that pattern that we talked about is much more nascent, much more obvious. And so, if we see that pattern, then that tells us perhaps we've missed something, that there's something else going on and usually what that something else going on is that we've modeled a nonlinear relationship as a linear one. Right?

So we have a curve, we try to fit a line to that curve, and as a result of that, these residuals or differences are a function of that underlying attribute. So let's look at a multiple regression example, so refocus on those attributes. Right? So if have a multiple regression example where we have two attributes, and we run our regression and generate that prediction and look at those residuals, initially as a function of our predicted Y, we see that there's a pattern, right? My ability to make an estimate is different for different values of whatever I'm trying to predict. Right?

So that tells us that there's some potential nonlinearities happening. It doesn't tell us what's driving those nonlinearities. So we dig deeper. And look at those residuals now as a function of the individual attributes. And so, for our first attribute, when we look at those residuals, they just seem to be sort of scattered loosely, in and around zero, for different values of the attribute. Right? We don't necessarily see any relationship between the residuals and the attribute itself.

Whereas, we look at our other attributes. We see that the residuals change. Right? They go from these positive numbers to negative numbers to positive numbers for different values of the attribute. And so, knowing the value of the attribute would tell us something about the residuals, i.e something about our ability to make an estimate for Y. So that informs us that there's some nonlinearity happening between that attribute and our dependent variable. So we would address that nonlinearity by doing a transform for this attribute but we would leave the other attribute alone because it is not displaying any nonlinearities with our predicted value.

Back to Table of Contents

Watch: Addressing Nonlinearities Using Transforms

You should always be prepared to consider where nonlinear relationships may exist in your data. In order to identify nonlinearities, you will need to be willing to interrogate the results of your initial regression and probe the data with residual plots. This is something that should become more natural with practice.

In this video, Professor Anderson introduces an example in which he demonstrates how to recognize and address a nonlinearity in an attribute. To address the nonlinearity, he uses a mathematical transform to modify the data in Excel.

Transcript

So as part of our regression modeling process, you're going to have to be able to detect nonlinearities, and then address them before you can proceed with your regression analysis. So let's go through an example. Megan has been recently hired at a financial services firm. And through casual discussions with other employees, she's come to the conclusion that it appears that similarly qualified females are getting paid substantively less than their male counterparts.

So, she has gone about collecting some data on starting salaries, is in the process of preparing a report that she can present to management around this gender inequity in starting salaries. So, part of our first step in our analysis is to sort of discover these relationships. So, we look at a scatter plot of starting salaries, and we look at those starting salaries for males, and females, as a function of time.

All right, so, if I was hired 10 years ago, we would expect to have a lower starting salary than if I was hired more recently. So now we see that relationship, we're going to proceed to building our regression models, we run a regression with two attributes: gender, and time. That resulting regression looks pretty informative, we have a decent R-squared. We have statistically significant coefficients, and the coefficient for gender is negative. Indicating that, on average, females are paid less than males.

As a quick diagnostic, we look at our residual plot and we specifically look at that residual plot was residuals as a function of time. And we see a pattern. Right? It looks like those residuals tend to be larger for smaller values of time, and then decrease and then start to increase again. Right?

So maybe there's some nonlinearities between time and starting salary. And that seems sort of logical. That yes, owing to inflation we would expect starting salaries to increase with time, but we wouldn't necessarily expect them to increase by a $1,000 each year as much as we'd expect them to increase by 5% each year. Right? So, because they're growing by this proportion versus growing by this constant level, we're going to have some nonlinearities between time and starting salaries. So let's address that. Let's do a transform. Let's take the natural log of salaries instead of salaries when we're running our regression.

So we revisit our regression now our dependent variable is the natural log of salaries. And again, we have a solid looking model because we've modeled this relationship better, when we look at our residual plots, now we see just sort of this randomness with those residuals as a function of time. Because we have no pattern in our residuals and we've done a better job at modeling the impact of time and salary, that translates to a higher R-squared, versus our mis-specified model. We have statistically significant coefficients, and now we can interpret the impact of gender.

So, because we, our dependent variable was the natural log of salaries and not salaries, we need to unravel that when addressing the impact of this coefficient. And so basically the impact of gender is going to be the exponential, or EXP in Excel, of that coefficient. So if we take the EXP of our parameter value, it turns out that that is our proportion of male salaries that females are making. And we see that that proportion is less than one, i.e. women are making less than those sort of equally qualified men. So, it's critical when we go through this process that we understand the dynamics between our attributes, and our dependent variables.

Back to Table of Contents

Watch: Modeling Interactions between Independent Variables

As you become comfortable with the procedures used to refine your model, you will be able to layer these procedures over one another, simultaneously considering nonlinearities and variable interactions. Your model will naturally become more complex. If you reason correctly and keep aligning your model with the reality of your situation, its predictive power should improve.

In this video, Professor Anderson extends a model with a nonlinear attribute to include an additional attribute that interacts with that nonlinear attribute. Using the refined model that results, he is able to draw further conclusions from the data available. This process of adding complexity to an existing model is one that should occur naturally as you create an initial regression model and then examine your scenario more deeply.

Transcript

So, when we run multiple regressions, we're looking at the impacts of attributes upon our dependent variable. And so, we could think about this as the impact of an attribute upon the dependent variable while controlling for the impacts of the other attributes. But sometimes, you want to go beyond just controlling for the other attributes. Sometimes, you want to look at the potential interactions between the attributes upon that dependent variable. So, let's think about our gender discrimination example, where originally we were looking at, is there a fixed difference between starting salaries for males versus females?

Let's extend that example. Instead of just looking at the differences between salaries themselves, but also are those differences between salaries growing with time? So was the salary itself different and are annual raises different? Right? So we're looking at the impact of gender and time, but also this interaction between gender and time. Is the impact of time different for males versus females? So if we didn't model the interaction between gender and time, in essence we have these two parallel lines. When we think about salaries, we have one series for

males and one series for females.

And so, we have this fixed difference between salaries, between our two genders, as a function of time. But now, if we think that males are potentially getting higher raises than females, then as time increases, then basically, the salaries between males and females are diverging. Right? So, there's just this fixed difference, yes that males are getting paid more than females, but they're getting high raises as well over time. And so how do we model that in our multiple regression world? So in essence what we have to do is we have to create a new variable, and that new variable is the product of gender and time. Right? So remember that we have coded gender as a one and zero. And, so, when we code that gender as a zero, then when that zero is multiplied by time, that is still a zero. So we have this variable, which is going to take on the value of time when gender is a one. And it will take on a value of zero when gender is a zero.

And so now we can run a regression where our dependent variable is going to be the natural log of salary, because again, we're hypothesizing these nonlinear relationships, between money and time. So the natural log of salary is going to be a function of gender. It's going to be a function of time, and it's also going to be a function of gender times time. So we run our regression, and we have statistically significant results. We have a decent R-squared, we have P-values for our coefficients which are all less than .05, so in statistically speaking our regression makes sense, and now we can actually interpret those coefficients. So we realize that this, we're going to have this fixed effect for gender. Right? So E to our gender co-efficient tells us the fixed difference between male and female salaries.

But then, if we think about the impact of time, we're going to have these two coefficients, which involve time, one which is just a function of our time variable, and one which is a function of our gender times time variable. And so, when gender is equal to a zero, then the only impact of time is our initial time variable. But when gender is equal to a one, then we have these two coefficients which are impacting time. And so, we have to add those together when we look at the impact of time. So now time is going to be a function of one coefficient for gender zero and it's going to be a function of these two coefficients for gender one. We take

each of to those coefficients.

Now we see that E to our gender zero value and then E to our gender one value is going to have different values, and we can see the difference between those raises for males versus females. If we take, given that our time is in months, if we take that coefficient to the power of 12, or each of that coefficient to the power of 12, we can transform that coefficient to an annual raise. And we'll see now that males are getting A, a higher salary and then B, are getting elevated annual raises. And so, we have sort of looked at these impacts of gender and time together versus them in isolation.

Back to Table of Contents

Watch: Considering Missing Variable Bias

Regression models are only as good as the data on which they are based. Sometimes there will be attributes for which you believe there is an association but you will lack the data needed to include them in your model. To the extent that you can, you should document these gaps in your model and describe the potential significance of the omission.

In this video, Professor Anderson discusses a hypothetical example in which a variable that is missing from the regression model might have a significant effect on the performance of the model.

Transcript

Part of our regression-building process is to map out the relationships, or the potential relationships, between all the factors that may be impacting our dependent variable. Then we would go out and collect information on all those factors, and run our regression to quantify the relationships between those attributes and our dependent variable.

More common though, is, you know, even though we may have that understanding of all those potential relationships, we can only collect information on a subset of those attributes. Or we may actually start with a data set, and that data set only includes a subset of the attributes, which are impacting our variable of interest. We're still going to run our regression, even though we only have a subset of the attributes, which may be impacting our variable of concern. We need to be cognizant, though, or aware, of the impact of exclusion of some of those variables.

So, let's look at an online retailer. This online retailer is trying to measure demand. They're trying to look at the key drivers of command. He has this strong intuition that price and user reviews are the two main drivers of demand. So, he goes out and collects some data, unfortunately he's unable to collect data on user reviews. He's only able to collect data on price. So he runs some regressions and determines the impact of price upon demand.

So, that's great, but he has to be aware and also acknowledge the impact

of exclusion of these user reviews. So, given that we sort of have this feeling that price and user reviews are probably related. Right? So if I have higher reviews, I probably price higher. So they're probably related. Then exclusion of user reviews means that the price variable is trying to do double lifting of the impact of price and user reviews. So, when we communicate our regression results, we have to acknowledge that we've excluded some of these key attributes. These attributes that we've excluded are potentially related or correlated to some of these attributes that we've included and that is going to bias, or impact, the coefficients of those parameters that we've included in our model.

Back to Table of Contents

Course Project, Part Three—Generate a Revised Regression Equation

You've seen how careful scrutiny of the intercept and coefficients can help you refine a regression. In this part of the course project, you will apply these techniques to the regression model you created in part two of the course project. When you complete this project part, you should have a refined regression model. *Completion of all parts of this project is a course requirement.*

Instructions:

1. Open your saved course project document. (If needed, download it again now.)
2. Complete Part Three.
3. Save your work.
4. You will submit your completed project at the end of the course for grading and credit.

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrap-up: Assessing and Validating Your Model

You began this module with an unrefined regression model based on a few key attributes. You progressed to a model with a higher fidelity and improved predictive power. The validation your model underwent involved examining the relationships between independent attributes for possible multicolinearity, the development of synthetic variables where appropriate, and adjustments for nonlinearities in the relationship between independent and dependent variables. Throughout the refinement process, you applied common sense about your situation based on your experience with the variables and operating parameters of the situation.

Congratulations. You now have a solid, validated regression model that works...on paper. In the next module, you will put your model to the test. Because no amount of theoretical performance is as meaningful as actual performance, you will want to see your model in action. In this last module of the course, you will complete the predictive analysis cycle.

Back to Table of Contents

Module 4: Applying the Predictive Analytics Framework

Back to Table of Contents

Module Introduction: Applying the Predictive Analytics Framework

No regression model is ever complete, or perfect. You can have a very good model, and you should strive for a degree of complexity and fidelity that meets your predictive or decision-making needs. Your ability—and that of your associates—to trust your regression model will be improved significantly the first time it performs well in a real situation.

In this module, you will test your regression model using a holdout sample. You will also consider how you might adjust your approach to regression if you are examining categorical variables or creating categorical variables within your quantitative data.

Back to Table of Contents

Watch: Using Your Regression Model

Now that you have created, refined, and validated your regression model, you will want to put it to use. What good is the model, otherwise? Perhaps a more pertinent question is, do you understand how the model you've created can be useful to you? Of course you should have begun the process with a clear idea of how you intended to use your regression model. But you've probably learned quite a bit about your problem along the way, and it's worth revisiting the initial premise to see if your target has changed or been refined.

In this video, Professor Anderson discusses the two broad purposes for which you may use your regression model. Stop to reconsider your model in the context of these two purposes. Are there useful opportunities that you didn't consider at the outset? He introduces the concept of a holdout sample as a means to apply a final validation test to your model, and he describes broadly how holdout samples work.

Transcript

So regression is a key part of our predictive analytics framework. We can basically use regression for two broad purposes. Right? We're going to take these attributes and look for relationships of those attributes with our dependent variable. One, to reduce uncertainty in that dependent variable. Looking for the impact of those attributes upon that dependent variable. So what's the impact of the number of pieces, or the number of steps in the completion time for a job to refine that estimate for the completion time.

Or, we could use that regression model and really focus on the coefficients for those attributes. What is the impact of the gender upon starting salary? I am not really trying to predict starting salary, I just want to know how much less females make versus males. And so, we sort of have this sort of process that we go through in regression. You know, we're starting with our ideas, looking for associations, checking for linearity, building models, making sure those models are valid both statistically and logically. And then, maybe circling back if we discover

some nonlinearities.

But at the end of the day, that regression is based on a sample. And we're really trying to take that model from that sample and generalize to the population. Make some prediction or some estimate of impacts about cases outside of our sample. And so, we have to ask ourselves how well does our model do at making these predictions outside of our sample? What is its predictive capabilities? And so, there's no real way to sort of test that without applying our model across other samples. And so, typically we do this through holdout samples. Right?

I have a sample of data I've collected. I am going to set aside 10 or 20% of that data. On the other remaining 80% or 90%, I'm going to go through my regression process and build some models. When I'm done, I'm going to take those, that regression equation on those predictions and apply it to the remaining 10% or 20% that I've set aside. And I'm going to see how well my model does on this holdout sample. If it's consistent across those samples and doing a good job on that holdout sample, I have more comfort in the generalizations of my model outside of this sample to the overall population in general.

Back to Table of Contents

Tool: Predictive Analytics Framework Diagram

Predictive Analytics Framework Diagram

The process of predictive analytics is cyclical. You begin with data for the situation or question of interest, develop a regression model, validate that model, and return to your starting point to test that model.

Keep this tool on hand to remind yourself of the steps of predictive analytics process.

Back to Table of Contents

Watch: Testing Your Model with a Holdout Sample

Testing your model with a holdout sample provides an extra degree of confidence in its usefulness. Depending on your situation, holdout sample testing may be a nice-to-have or a hard requirement for validation. Consider the stakes, the tolerance for error, and how broadly (and by whom) your model will be used.

In this video, Professor Anderson works through an example that illustrates in detail the testing of a regression model using a holdout sample.

Transcript

All right, so let's look at our predictive analytics framework in action. Easy Readers is a book club that direct markets to consumers. It has a large database and, at present, it's not necessarily uniquely targeting members of that database with offers. It's looking to send out books to consumers. Obviously, the sending out of those books incurs some cost. And if the consumers actually purchase those books, there's going to be a revenue stream.

In order to develop methods for more accurately targeting consumers, they conduct an experiment. They send out a book to 2,500 consumers. We're going to take a subset of that 2,500, say 2,000 of those, and we're going to build some regression models on those 2,000. And then we're going to test, or apply, that model to a holdout sample of 500 consumers to see how well that model performs. All right, so let's look at some data. So our sample has information on whether or not consumers purchased this new title. And then information across four attributes for each of those consumers, their gender, their total spend with us, the number of total purchases with us, as well as the number of purchases for similar young reader books.

We can look at the averages of those four attributes and get some insight that those attributes are different across our dependent variable, and may be useful in explaining that dependent variable. We can look at the

distribution of those attributes, and by that I mean we look at the distribution of total purchase spend in dollars, total number of purchases, and total number of young reader purchases, but we have distributions for both those who purchased our new title and those that didn't purchase our new title, providing further support for those attributes being part of our regression model. We run our regression with those four attributes, we notice the potential for multicollinearity between the number of purchases and the number of young reader purchases.

So we drop the number of purchases, we re-run our regression, our R-squared does not change substantively with our modified regression and we have both logical and statistically significant coefficients. And so, now we can sort of use this regression model to help us better target consumers. To do that, we basically look at our sample, we take the attributes for each member of that sample and apply our regression equation to those attributes. We generate a regression score based upon those attributes. We could take our sample of 2,000 consumers and sort that regression equation from largest to smallest.

Again, largest being closer to one, smallest being closer to zero. And then, we could look at our sample and split it into quartiles. Right? So look at the top, the 25% with the highest score, the next 25% etc., across our four different quartiles. Once we have our sample subset in these quartiles, we can look at how many of people in that quartile purchased our new title. And given that number of purchases, if we had targeted that quartile, what would have been our profit as a function of our cost to mail out the book and the profit we net from the consumers keeping that book, or buying that book. Right?

So we have a sense here that targeting the top quartile is profitable, but targeting the other three quartiles is not profitable. So, now we're going to test our model on our holdout sample to see how well it performs sample and, in essence, is our model going to be reliable going to our general population of all our customers? And so, we take our holdout sample and, as with our regular sample, we apply our regression equation to the attributes for each of our 500 consumers in our sample. We sort them again and we look at the total number of purchases across each of those quartiles.

And again, we see that in our holdout the top quartile is profitable, the other three quartiles are not, and we can actually compare the profit from targeting this top quartile versus not targeting consumers and just sending out our mailer to everyone. If we did that, we would actually generate a substantive loss, and so here we have the gain from using out model. Right? That gain is the profit from targeting this quartile plus this loss that we can avoid from not just blanket sending out our mailing. And so, this performance in our holdout, being consistent with the sample that we built our regression on, provides evidence that our model will perform well outside of the sample that we use to create it.

Back to Table of Contents

Watch: Using Logistic Regression to Model Categorical Variables

So far in this course we have focused on regression models that involve quantitative variables. In situations where the dependent variable is categorical, you will need to modify your modeling approach in order to obtain useful results.

In this video, Professor Anderson gives an overview of logistic regression and discusses different situations that call for different types of logistic regression.

Transcript

Sometimes, when we're looking at a categorical dependent variable in our effort to find associations between that categorical variable and other variables, we need to adjust our approach to regression. For example, if we had a categorical dependent variable which had outcomes of zero and one, and had a quantitative independent variable. If we were to try and fit a line to that relationship, we'd quickly realize that that line is going to extend well beyond our two outcomes for that categorical dependent variable. And so, trying to fit a simple line to that perhaps is not the best approach. We could think of this also in the terms of correlations.

Our traditional correlation coefficient, or what sometimes referred to as our Pearson correlation coefficient, basically looks at the differences between two variables and their means, and how those two variables move relative to their individual means. So in that context, the mean of that variable is important. But when I have a categorical variable, I'm arbitrarily sort of associating numbers with those categories. So I might have three categories. I could code those three categories 1, 2, 3 or 1, 6, and 12. And so, the mean across those three categories is rather meaningless.

So when I look at correlations with categorical variables I need to use what's called a person Pearson correlation quafifient,. We we're looking at the rank of that categorical variable versus it's value relative to it's

mean. So if I have two categorical variables, right? So both a categorical dependent variable and a categorical independent variable. And I'm trying to look for associations between those two variables. We can look for those associations by simply focusing on proportions. So let's focus on two nationalities and their preference for red versus white wine. We see that both nationalities have the same preference for red wine and white wine.

So that would tell us that there's no association Between those nationalities and wine preference. Where if we looked at those same two nationalities, but now instead of looking at wine preference, we looked at beverage preference, so beer versus wine, we noticed that There is a difference in those proportions and so knowing which nationality you are is informative of whether or not you would prefer beer versus wine. So we see an association between nationality and beverage preference. Where was we didn't see one between nationality and wine Preference.

Now, if we have a categorical dependent variable and a quantitative or continuous independent variable, so and we're looking at trying to use regression to help us understand that association. So our line is not logical, what we really want to focus on is fitting our curve to that dependent variable such that the curve doesn't extend beyond the categories and switches quickly from one category to the other. We refer to this type of regression as logistic regression versus linear regression. Right?

So, we have, it's still regression, it's just slightly different flavor. And so, depending upon the nature of our dependent variable, we will use different forms of logistic regression. If we simply have two categories, so, a binary one-zero or yes-no dependent variable, we will use what's called binary logistic regression. But now if I have three or more categories, we have to look at if the categories themselves are informative. Right?

So we have two types of categorical variables and we have three or more categories. We can think of those as nominal and ordinal. Where ordinal categorical variables have a natural order. Right? So you can think of those as small, medium and large. So if I've coded small, medium and large, one, two and three, then the order of those categories is

informative.

Whereas, if I'm just looking at three choices, A, B and C, that I've coded one, two and three, then there is no natural sequence to those categories, so we refer to those as nominal and we would run a nominal logistic regression. All right? So, we're still going to approach this in our regression framework, we're just going to use slightly different tools. But for all intents and purposes, how we sort of use those tools and how we infer the strength of those tools is going to be consistent with how we tackled things in our linear regression world.

Back to Table of Contents

Watch: Segmenting by Creating Artificial Categories

Sometimes it will be useful to create groupings or ranges within a quantitative variable in order to create a categorical variable with distinct, finite categories. Segmenting a market is a common situation in which you might code quantitative data in terms of a categorical variable. The choice to create an artificial categorical variable should follow from a specific purpose that you cannot fulfill by looking at the data for an attribute (or for your dependent variable) as continuous variable.

In this video, Professor Anderson discusses the rationale for creating artificial categories within a data set. He considers a scenario in which artificial categories are used to define a problem more usefully than would be possible treating the data qualitatively.

Transcript

So in practice, many natural categories exist within the data we collect. Those categories might be demographic. They may be purchase versus no purchase, or purchase product A versus purchase product B. But sometimes, our data has quantitative variables, and we may want to artificially create categories based upon, or derived from, those quantitative variables.

So think of a firm that's tracking customer purchases, or total expenditures with the firm, and we realize that we're really focused on is profitable customers versus unprofitable customers. So we take this total expenditure, quantitative variable, and we separate it into those that spend more than a certain level as being profitable, and those that spend less being unprofitable. And so, we've created this two arbitrary categories and now we might go back in look for associations, for, across other attributes that helps us inform who might be a profitable consumer and which consumers we want to develop or continue relationships with. Sometimes, we'll create these categories from more than one variable and those variables may be quantitative.

So, let's look at a firm that is selling a series of products and we have

sales of various products. We can look at customers who are purchasing these low price, low quality products. We have another subset of customers who are making purchases at a higher prices for higher quality products. And then, this third group of customers that are kind of clustered around these higher prices and low to moderate quality. So we have these two quantitative variables and we look, or basically we focus, on these three segments or these three groups of consumers. And now that we've separated consumers into these three segments, we can look for associations, which may inform us which segment a consumer falls into. Right?

So we may use logistic regression now with these three segments as our dependent variables, focusing on a series of other attributes to inform us of which segment our customers may lie in and how we might target them with the appropriate marketing actions. So it's critical for us to sort of understand what happens when we have natural categories or if we artificially create those categories, and how we want to look for relationships across those categories and other attributes.

Back to Table of Contents

Course Project, Part Four—Validate Your Model

In the first three parts of your course project, you worked through the mechanics of accumulating evidence in support of your model. In this part of the course project, you synthesize your findings in a concise and minimally technical argument for the validity of your model. *Completion of all parts of this project is a course requirement.*

Instructions:

1. Open your saved course project document. (If needed, download it again now.)
2. Complete Part Four.
3. Save your work.
4. Once you've finished, review the entire document, making any final additions or revisions, and then **submit it for instructor review using the Submit Assignment button on this page.**

Do not hesitate to contact your instructor if you have any questions about the project. You will add to this document as the course proceeds and will submit it to the course instructor at the end of the course.

Before you begin:

Before starting your work, please review the **rubric** (a list of evaluative criteria) for this assignment. Also review eCornell's policy regarding **plagiarism** (the presentation of someone else's work as your own without source credit).

Back to Table of Contents

Module Wrapup: Applying the Predictive Analytics Framework

In this module, you considered whether your goals could be met by introducing a categorical variable to your regression analysis. If appropriate, you used a holdout sample to test your regression model.

You have completed a broad survey of the predictive analytics framework. To be sure, there is more to learn about tools and specific techniques to improve the results of predictive modeling. But you should be leaving this course with a more complete understanding of the mindset and approaches that result in effective models for prediction and decision making. Additional techniques can be understood within the context of this framework and added to your repertoire of predictive analytics skills. So long as you keep your problem context in mind and continue to revisit and refine your regression, you should be able to develop a powerful predictive model.

Back to Table of Contents

Read: Thank You and Farewell

Chris Anderson
Associate Professor
School of Hotel Administration
Cornell University

Congratulations on completing *Using Predictive Data Analysis*. I hope your work in this course has left you better prepared to make confident decisions and predictions based on statistical methods and your understanding of your domain of expertise.

From all of us at Cornell University and eCornell, thank you for participating in this course.

Sincerely,

Chris Anderson

Back to Table of Contents

1. Excel Step-by-Step Instructions

Back to Table of Contents

Excel Step-by-Step Instructions

Use these guides to help you complete data analysis tasks in Excel. Instructions are provided for Excel for Windows and Excel for Mac.

Back to Table of Contents