

Introduction

The Data Analysis Iteration

Six Types of Questions

Characteristics of a Good Question

Exploratory Data Analysis

Using Models to Explore Your Data

Exploratory Data Analysis: When to Stop Inference

Formal Modeling

Inference vs. Prediction: Implications for Modeling Strategy

Video: Inference vs. Prediction 12 min

Reading: Inference vs. Prediction 10 min

Quiz: Inference 5 questions

Quiz: Formal Modeling, Inference vs. Prediction 5 questions

Interpretation of Results

Communication

Post-Course Survey

Understanding whether you're answering an inferential question versus a prediction question is an important concept because the type of question you're answering can greatly influence the modeling strategy you pursue. If you do not clearly understand which type of question you are asking, you may end up using the wrong type of modeling approach and ultimately make the wrong conclusions from your data. The purpose of this reading is to show you what can happen when you confuse one question for another.

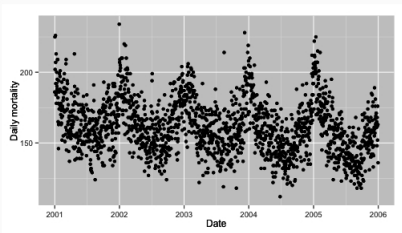
The key things to remember are

- For **inferential questions** the goal is typically to estimate an association between a predictor of interest and the outcome. There is usually only a handful of predictors of interest (or even just one), however there are typically many potential confounding variables to consider. They key goal of modeling is to estimate an association while making sure you appropriately adjust for any potential confounders. Often, sensitivity analyses are conducted to see if associations of interest are robust to different sets of confounders.
- For **prediction questions** the goal is to identify a model that *best predicts* the outcome. Typically we do not place any *a priori* importance on the predictors, so long as they are good at predicting the outcome. There is no notion of "confounder" or "predictors of interest" because all predictors are potentially useful for predicting the outcome. Also, we often do not care about "how the model works" or telling a detailed story about the predictors. The key goal is to develop a model with good prediction skill and to estimate a reasonable error rate from the data.

Air Pollution and Mortality in New York City

The following example shows how different types of questions and corresponding modeling approaches can lead to different conclusions. The example uses air pollution and mortality data for New York City. The data were originally used as part of the National Morbidity, Mortality, and Air Pollution Study (NMMAPS).

Below is a plot of the daily mortality from all causes for the years 2001–2005.



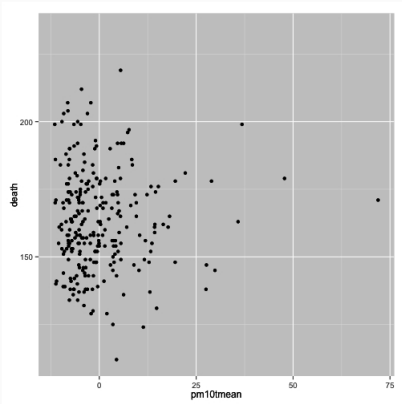
And here is a plot of 24-hour average levels of particulate matter with aerodynamic diameter less than or equal to 10 microns (PM10).

Note that there are many fewer points on the plot above than there were on the plot of the mortality data. This is because PM10 is not measured everyday. Also note that there are negative values in the PM10 plot—this is because the PM10 data were mean-subtracted. In general, negative values of PM10 are not possible.

Inferring an Association

The first approach we will take will be to ask "Is there an association between daily 24-hour average PM10 levels and daily mortality?" This is an inferential question and we are attempting to estimate an association. In addition, for this question, we know there are a number of potential confounders that we will have to deal with.

Let's take a look at the bivariate association between PM10 and mortality. Here is a scatterplot of the two variables.



There doesn't appear to be much going on there, and a simple linear regression model of the log of daily mortality and PM10 seems to confirm that.

		Estimate	Std. Error	t value	Pr(> t)
1	(Intercept)	5.08864300354	0.0060353779	733.75138151	0.0000000
2	pm10tmean	0.00004033446	0.0006913941	0.05833786	0.9535247

In the table of coefficients above, the coefficient for pm10tmean is quite small and its standard error is relatively large. Effectively, this estimate of the association is zero.

However, we know quite a bit about both PM10 and daily mortality, and one things we do know is that *season* plays a large role in both variables. In particular, we know that mortality tends to be higher in the winter and lower in the summer. PM10 tends to show the reverse pattern, being higher in the summer and lower in the winter. Because season is related to *both* PM10 and mortality, it is a good candidate for a confounder and it would make sense to adjust for it in the model.

Here are the results for a second model, which includes both PM10 and season. Season is included as an indicator variable with 4 levels.

	Estimate	Std. Error	t value	Pr(> t)
1				
2 (Intercept)	5.166484285	0.0112629532	458.714886	0.000000e+00
3 seasonQ2	-0.109271301	0.0166902948	-6.546996	3.289291e-10
4 seasonQ3	-0.155803242	0.0169729148	-9.161847	1.736346e-17
5 seasonQ4	-0.068317619	0.0167189714	-3.687735	3.716291e-04
6 pm10tmean	0.001499111	0.0006156902	2.434847	1.558453e-02

Notice now that the pm10tmean coefficient is quite a bit larger than before and its t value is large, suggesting a strong association. How is this possible?

It turns out that we have a classic example of [Simpson's Paradox](#) here. The overall relationship between P10 and mortality is null, but when we account for the seasonal variation in both mortality and PM10, the association is positive. The surprising result comes from the opposite ways in which season is related to mortality and PM10.

So far we have accounted for season, but there are other potential confounders. In particular, weather variables, such as temperature and dew point temperature, are also both related to PM10 formation and mortality.

In the following model we include temperature (tmpd) and dew point temperature (dptp). We also include the date variable in case there are any long-term trends that need to be accounted for.

	Estimate	Std. Error	t value	Pr(> t)
1				
2 (Intercept)	5.62066568788	0.16471183741	34.1242365	1.851690e-96
3 date	-0.000802984198	0.00001315212	-2.2689856	2.431521e-02
4 seasonQ2	-0.05805970053	0.02299356287	-2.5250415	1.218288e-02
5 seasonQ3	-0.07655519887	0.02904104658	-2.6361033	8.906912e-03
6 seasonQ4	-0.01546943985	0.01832712585	-1.7213252	8.641910e-02
7 tmpd	-0.00295931276	0.00128835065	-2.2969777	2.244054e-02
8 dptp	0.00068342228	0.00103489541	0.6603781	5.096144e-01
9 pm10tmean	0.00237049992	0.00065856022	3.5995189	3.837886e-04
10				

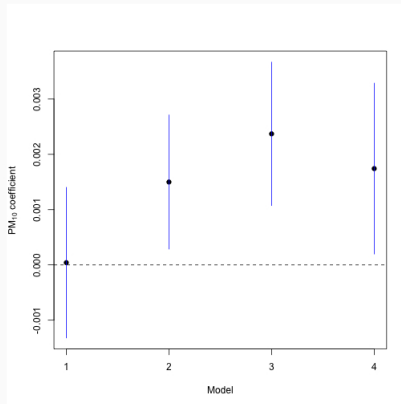
Notice that the pm10tmean coefficient is even bigger than it was in the previous model. There appears to still be an association between PM10 and mortality. The effect size is small, but we will discuss that later.

Finally, another class of potential confounders includes other pollutants. Before we place blame on PM10 as a harmful pollutant, it's important that we examine whether there might be another pollutant that can explain what we're observing. NO2 is a good candidate because it share some of the same sources as PM10 and is known to be related to mortality. Let's see what happens when we include that in the model.

	Estimate	Std. Error	t value	Pr(> t)
1				
2 (Intercept)	5.61378684085	0.16440280471	34.1465345	2.548704e-96
3 date	-0.000802973484	0.00001312231	-2.2659756	2.430503e-02
4 seasonQ2	-0.05143935218	0.02338034983	-2.2001105	2.871069e-02
5 seasonQ3	-0.06569205605	0.02990520457	-2.1966764	2.895825e-02
6 seasonQ4	-0.02758381423	0.01849165119	-1.4873659	1.381739e-01
7 tmpd	-0.00296833498	0.00128542535	-2.3092239	2.174371e-02
8 dptp	0.00070366996	0.00103262057	0.6808599	4.965877e-01
9 no2tmean	0.00126556418	0.00086229169	1.4676753	1.434444e-01
10 pm10tmean	0.00174189857	0.00078432327	2.2288937	2.725117e-02

Notice in the table of coefficients that the no2tmean coefficient is similar in magnitude to the pm10tmean coefficient, although its t value is not as large. The pm10tmean coefficient appears to be statistically significant, but it is somewhat smaller in magnitude now.

Below is a plot of the PM10 coefficient from all four of the models that we tried.

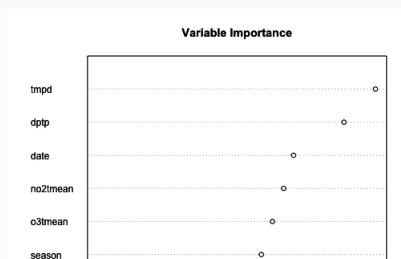


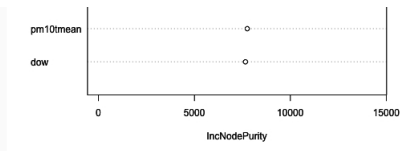
With the exception of Model 1, which did not account for any potential confounders, there appears to be a positive association between PM10 and mortality across Models 2-4. What this means and what we should do about it depends on what our ultimate goal is and we do not discuss that in detail here. It's notable that the effect size is generally small, especially compared to some of the other predictors in the model. However, it's also worth noting that presumably, everyone in New York City breathes, and so a small effect could have a large impact.

Predicting the Outcome

Another strategy we could have taken is to ask "What best predicts mortality in New York City?" This is clearly a prediction question and we can use the data on hand to build a model. Here, we will use the random forests modeling strategy, which is a machine learning approach that performs well when there are a large number of predictors. One type of output we can obtain from the random forest procedure is a measure of variable importance. Roughly speaking, this measure indicates how important a given variable is to improving the prediction skill of the model.

Below is a variable importance plot, which is obtained after fitting a random forest model. Larger values on the x-axis indicate greater importance.





Notice that the variable `pm10tmean` comes near the bottom of the list in terms of importance. That is because it does not contribute much to predicting the outcome, mortality. Recall in the previous section that the effect size appeared to be small, meaning that it didn't really explain much variability in mortality. Predictors like temperature and dew point temperature are more useful as predictors of daily mortality. Even `NO2` is a better predictor than `PM10`.

However, just because `PM10` is not a strong predictor of mortality doesn't mean that it does not have a relevant association with mortality. Given the tradeoffs that have to be made when developing a prediction model, `PM10` is not high on the list of predictors that we would include—we simply cannot include every predictor.

Summary

In any data analysis, you want to ask yourself "Am I asking an inferential question or a prediction question?" This should be cleared up before any data are analyzed, as the answer to the question can guide the entire modeling strategy. In the example here, if we had decided on a prediction approach, we might have erroneously thought that `PM10` was not relevant to mortality. However, the inferential approach suggested a statistically significant association with mortality. Framing the question right, and applying the appropriate modeling strategy, can play a large role in the kinds of conclusions you draw from the data.

NOTE: Part of this reading is taken from [The Art of Data Science](#) by Peng & Matsui.

Mark as completed