Chapter 1 and 2: Introduction and Statistical Learning

Andrew Andrade

2/19/2017

Contents

What is Statistical Learning?

Let's stat off with an analogy:

Probability is starting with an animal and figuring out what footprints it will make.

Statistics is seeing a footprint and guessing the animal.

Statistical learning is the study of seeing footprints, guessing animals, verifying the guesses and understanding why they made those types of footprints.

This guide will focus just on statistical learning: modeling real world phenomena both predicting "what" is going to happen based on past observations, and using inference to explain the "how" and the "why". Let's start with a toy example.

Predicting Marks

For a simple example, let's say we want to predict what mark someone would get on a test taken on a specific course. To do this we used what data was available to us: the number people's visits to the course website, how long they spend on the site and their mark on the test. How do we represent this problem?

The measured value we are trying to estimate in this case is the mark on the test (a continuous value from 0-100%). We are using 2 types of recorded measurements/inputs: a number of visits to the website (counting number) and the duration of time studying. (continuous value). Now we have 100 students in our class who are taking the test. There is also going to be some error in the prediction. For example, we won't be able to estimate their mark perfected by just measuring the logins. Maybe people who are logged onto the site might not have been studying the full time, for example watching TV in the background. If the measurements don't represent reality, there is a flaw in the measurement which leads to error. Who knows, maybe the system running the course website might have a bug and incorrectly counts logins or time spent browsing. Potentially more important, there are many other factors which would impact their mark which is not (and can not be) potentially measured (their skill level, how much sleep they got, how much they dislike their teacher etc.) The point is, there is going to be some error irrespective on what data you collect, how you collect it or what function you choose.

Now, let's formulate this information in math notation.

We want to predict Y using a function which will call f() and inputs X.

$$Y = f(X) + \epsilon$$

In this representation of the word, the thing we are trying to predict (Y) is a function (f()) of some measured variables (X) and some random error ϵ. In our example, this random error can come from many sources but are mainly (1) flaws in measurement, (2) some other factors influencing the result which are not measured

Now we can think of X as the matrix of observations, with n as the number of observations, and p are the number of types of inputs/measurements (commonly known as feature variable or features). n = 100 in our example since we have 100 students, and p =2 since we have 2 predictors/inputs/features (number of logins, and time spent studying).

For readers who are unfamiliar with matrices, it is useful to visualize X as a spreadsheet of numbers with n rows and p columns.

$$X = \begin{bmatrix} Feature_1 & Feature_2 & \dots & Feature_n \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ ... & ... & ... & ... \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{bmatrix}$$

Y is the response can be thought as a vector (a matrix with 1 column) with a measured response for each measurement.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$f()$ is f is some fixed but unknown function which maps our measured inputs (X1, . . . , Xp) to the response (Y) and ϵ is a random error term, which is independent of inputs X and has mean zero. Having a mean zero means that the average error will be zero, so there will be an error in either direction. In this case, the "best" model selected will sometimes guess someones mark too high or too low, but on average the guess will be correct.

In the example of predicting a score on the test:

$$X = \begin{bmatrix} \text{Time Logged In (Hours)} & \text{Number of Log-ins} \\ 2.74 & 5 \\ 4.07 & 5 \\ ... & ... \\ 1.92 & 2 \end{bmatrix}$$

$$Y = \begin{bmatrix} Mark(\%) \\ 71.0 \\ 79.3 \\ \vdots \\ 63.9 \end{bmatrix}$$

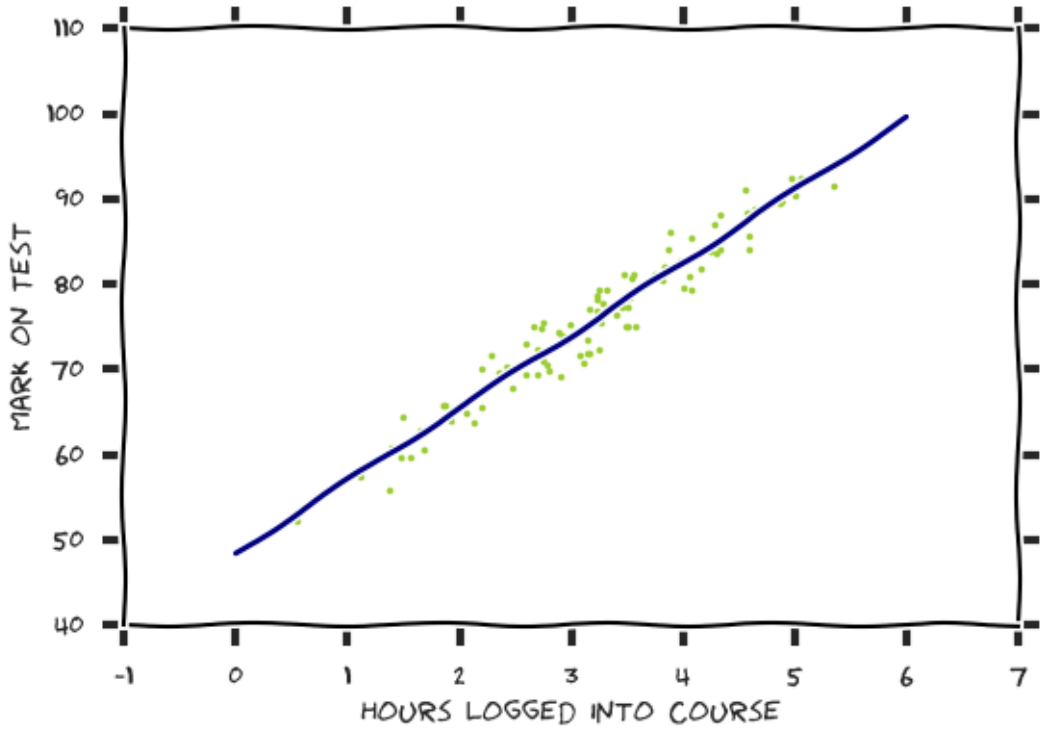

Figure 1:

Let's finish our example with an estimated function which presents how well people do on a test. Let's say that after plotting the data, we realize that there is a linear relationship between the number of hours students are logged into the course website, and the mark they get on a test.

Well, we can draw a line to fit the data, and use this to both predict the mark of new students and even understand the relationship between hours logged into the course website and student's performance on a test. In reality though, the model of the world is not so simple and data in real life is dirty. Let's look at the same example problem, but a different set of data:

In this data there are an obvious group of outliers in the top left: they have very high marks, yet they have only logged into the system for a short period of time. Outliers sometimes represent gold in a dataset, and other times trash. Maybe the system is logging the time spent online is incorrect, maybe these groups of students are cheaters or maybe truthfully wern't longed in for long yet did very well. As you will soon learn there are many robust statistical methods you can use to fit the data. For example, you can exclude the outliers and fit a model based on the inliers.

Using stastical learning, we can estimate an understanding of the world (f()) based on the observations we make. We will soon learn how to do this but first: Why estimate f()?

Why estimate f( )?

We want to estimate f() for two reasons: (1) Prediction or (2) Inference

1. Prediction

Since we don't have information on the full population of students, we have to estimate what f() looks like based on our observations. Our goal is to estimate the mark that *anyone* would get on the test, given the

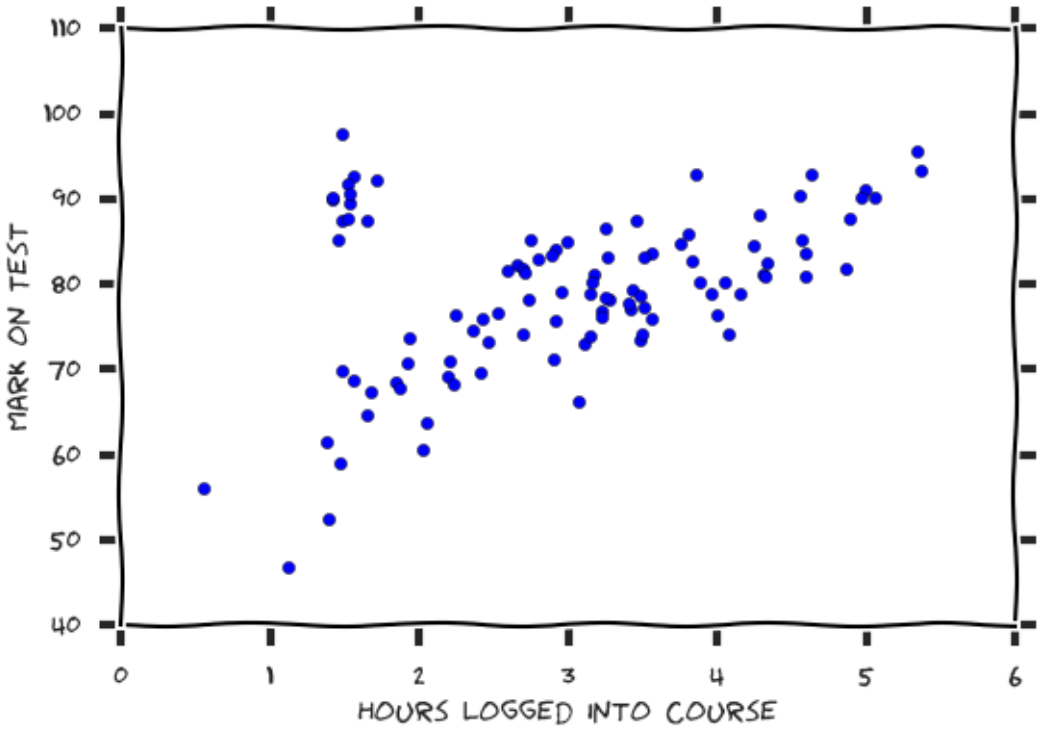

Figure 2:

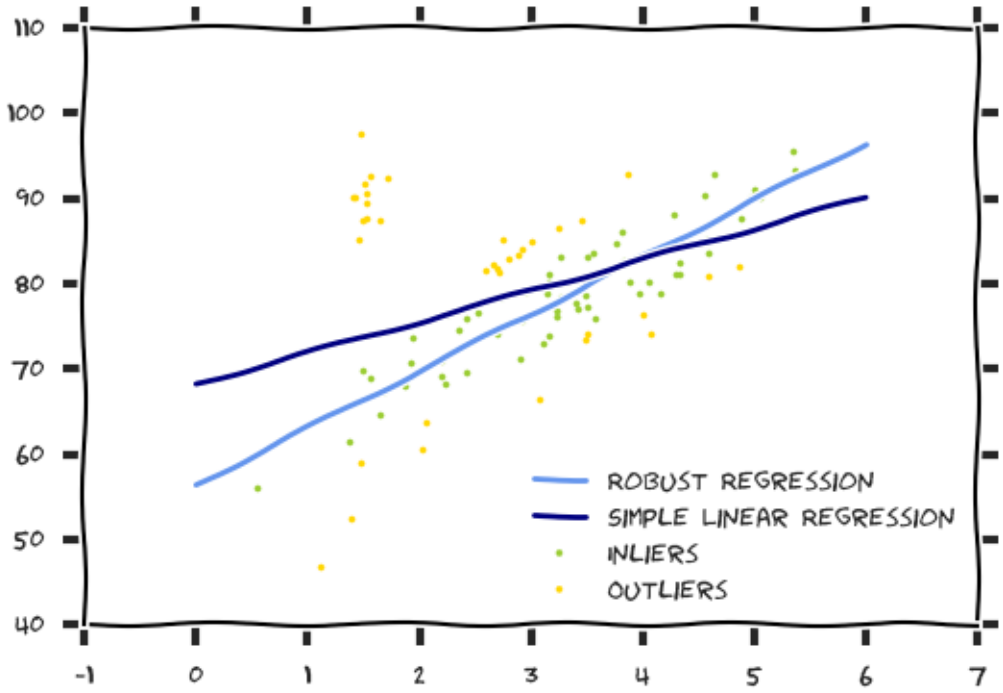

Figure 3:

information from only 10 students took the class. You can imagine, in the real world, we would want a larger sample size (a large number of n observations) and some more features/predictors (p) which better represent what their mark will be on the test.

Since the function we are estimating is not the actual full population of people who could take the class and do (we don't have data on future enrollment and performance), we use a ˆ to indicate that it's a prediction or estimation of the population.

$$\hat{Y} = \hat{f}(X)$$

$\hat{f}$ is estimate for $f()$ and $\hat{Y}$ is the prediction for Y.

Generally, most people treat $\hat{f}()$ as a black box and don't care the exact form of $\hat{f}()$ so long as it accurately predicts Y. Determining a "better" (more accurate, more precise, etc.) $\hat{Y}$ is the focus of prediction while understanding the mechanisms which case the response and quantifying them is inference (which we learn about in the Inference section).

Measuring "goodness"

"Goodness" and "better" are relative terms, so the first thing to do in prediction after determining what is being measured as the response, our inputs or feature variables (commonly known as features)

For numerical prediction (called regression), to test how well the model fits observations, we can take the difference between what we measured actually happened (Y) and the predicted response ($\hat{Y}$):

$$Error = Y - \hat{Y}$$

This isn't the best way to measure incorrectness. Here is an example, let's say we guessed that a student would get 100% on a test but they got 90%. We were wrong by 10% (+10% wrong according to the error formula). Similarly, we guessed a student would get 50% but they actually got 60% (-10% wrong according to the error formula). In this case we too were wrong by 10%, yet the average error would be 0% ($frac{-10\% + 10\%}{2} = 0$). This can easily be interpreted as no error since a perfect model would have an average error of 0. Yikes!

To fix this we want to get the absolute error (make the error positive whenever we are wrong). Another thing we would want to do is to penalize larger errors. For example, estimating a mark of 50% when the student gets a 60% (difference of 10%) is worse than estimating 55% (difference of 5%). Let's call the function that tells us how "bad" our model estimation is the expected value function (E()). A simple way to make the expected value of the error or difference in our prediction and actual measurement positive, and penalize incorrectness is to take the square of the error (also called quadratic loss).

Let's describe that in math speak. Now since we know the formula for $\hat{Y} = \hat{f}(X)$ and the general form of our observations $Y = f(X) + \epsilon$, we can sub it in to get the follow formula:

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$$

$$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + Var(\epsilon)$$

$E(Y - \hat{Y})^2$ represents the average of the squared difference between the predicted and actual value of Y. $Var(\epsilon)$ represents the variance associated with the error term ϵ. ϵ, while having a mean of 0, will have variance associated with is unavoidable.

The accuracy of our estimation model $\hat{Y}$ depends on the reducible error $[f(X) - \hat{f}(X)]^2$ and the irreducible error $Var(\epsilon)$.

The reducible error is reduced by choosing an appropriate model $\hat{f}()$ while the irreducible error is a function of ϵ by definition is not reduced by choosing a model. The quantity ϵ exists because we live in a real world: it may contain unmeasured variables that are useful in predicting Y: since we don't measure them, we cannot use them for its prediction.

The focus of statistical learning is on techniques for estimating f() with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

2. Inference

The second thing which we are interested in is understanding how the response we are measuring (Y) is affected by the values of our inputs/features $(X_1, ...X_p)$ change. We want to estimate (f()), but the goal is not necessarily to make "great" predictions, rather understand how our output changes as a function of the input.

We want to answer:

1. Which predictors are associated with the response?
2. What is the relationship between the response and each predictor?
3. Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

For example in advertising:

1. Which media contribute to sales?
2. Which media generate the biggest boost in sales? or
3. How much increase in sales is associated with a given increase in TV advertising?

Finally, some modeling could be conducted both for prediction and inference. For example, in a real estate setting, one may seek to relate values of homes to inputs such as crime rate, zoning, distance from a river, air quality, schools, income level of community, the size of houses, and so forth. In this case, one might be interested in how the individual input variables affect the prices that is, how much extra will a house be worth if it has a view of the river? This is an inference problem. Alternatively, one may simply be interested in predicting the value of a home given its characteristics: is this house under- or over-valued? This is a prediction problem.

How Do We Estimate f ( ) ?

The first step in estimating is setting up a set of training data. These observations are called the training data because we will use these observations to train, or teach, our method how to estimate f(). Our goal is to apply a statistical learning method to the training data training data in order to estimate the unknown function f(). In other words, we want to find a function estimate $\hat{f}$ such that approximates the observed response $(Y \approx f(X))$ for any observation (X, Y). We want the model/function to be generalized. This means that our fundamental goal is to find a model which is generalized beyond the examples in the training set.

From Professor Pedro Domingos's A few useful things to know about machine learning:

This is because, no matter how much data we have, it is very unlikely that we will see those exact examples again in the future. The most common mistake among beginners is to test on the training data and have the illusion of success. If the chosen model is then tested on new data, it is often no better than random guessing. So, if you hire someone to estimate a model, be sure to keep some of the data to yourself and test the model they give you on it. Conversely, if you've been hired to build a model, set some of the data aside from the beginning, and only use it to test your chosen model at the very end, followed by learning your final model on the whole data.

We will learn the details of splitting training and test data, later, but it is very important that you ***never estimate models on the full dataset*** as this can lead to models which do not reflect reality. Since there

will always be an error, it is good to keep in mind that there will be an error in both the training and testing of the models. The important thing is that error on the training set and the testing set be similar. It is better to have a model which is correct 75% on both the training set and testing set than a model which is 100% in the training set and 60% on the test set.

The general approach to model estimation is:

1. Representation: the model must be represented in some formal language that the computer can handle.
2. Evaluation: an evaluation function (also called objective function or scoring function) is needed to distinguish "good" models from bad ones
3. Selection: Finally, we need a method to search among the models for the highest-scoring one based on the test data (held out during training).

Let's tackle representation first. Broadly speaking, most statistical learning methods for this task can be represented as either a parametric or non-parametric model.

Parametric Methods

Parametric methods involve a two-step model-based approach.

1. Functional form: We make an assumption about the functional form, or shape, of f().

The simpliest form of a model is a linear (described extensively in chapter 3):

$$f(X) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

This model is defined by a number of parameters ($\beta_0..\beta_P$). This means we only need p+1 coefficients to describe the model. In addition, we can use the parameters to get a better understanding of the effects of each of our features (Xp). Parametric models are generally simply its much easier to estimate a set of parameters fit an entirely arbitrary functions. The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f, then our estimate will be poor. We can try to address this problem by choosing flexible models that can fit many different possible functional forms for f. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as overfitting the data, which essentially means they follow the errors, or noise, too closely. These issues are discussed through out the book and this work.

2. After a model has been selected, we need a procedure that uses the training data to fit or train the model. In the case of the linear model, we need to estimate the parameters ($\beta_0, \beta_1 + ... + \beta_p$) such that the linear model is approximately equal to our measure output ($Y \approx \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$).

The most common approach to fitting the linear model is referred to as (ordinary) least squares, which we discuss in Chapter 3. However, least squares is one of many possible ways way to fit the linear model.

Parametric Model Example

For example, let's say we are statistical consultants working for a firm who is trying to improve sales by running an advertising campaign. In this case, we know historically how much money was spent on Radio and TV advertising, and what the sales were during those times.

Our data matrices look like the following:

$$X = \begin{bmatrix} Radio(\$M) & TV(\$M) \\ 37.8 & 230.1 \\ 39.3 & 44.5 \\ ... & ... \\ 8.6 & 232.1 \end{bmatrix}$$

$$Y = \begin{bmatrix} Sales(\$MM) \\ 22.1 \\ 10.4 \\ \vdots \\ 13.4 \end{bmatrix}$$

Now we want to estimate a model to determine the Sales based on Radio and TV advertising. A simple parametric model would fit a linear model to the data. In 2 dimensions we would fit a line, and in 3 dimensions (sales, radio, and TV) we would fit a linear plane.

This linear model would be in the form:

$$sales \approx \beta_0 + \beta_1 Radio + \beta_2 TV$$

$$\begin{bmatrix} Sales(\$MM) \\ 22.1 \\ 10.4 \\ \vdots \\ 13.4 \end{bmatrix} = \beta_0 + \beta_1 \begin{bmatrix} Radio(\$M) \\ 37.8 \\ 39.3 \\ \vdots \\ 8.6 \end{bmatrix} + \beta_2 \begin{bmatrix} TV(\$M) \\ 230.1 \\ 44.5 \\ \vdots \\ 232.1 \end{bmatrix}$$

Fitting it to actual data from the `Advertising` dataset we get the following:

A plane in 3 dimemions is like a piece of paper. It's hard to visualize what it looks like in 4 or more dimensions, but planes work the same way. Our goal in fitting a the lienar model is we want to orient the paper such that the paper is as "close" to as many points as possible"."Closeness" is defied by the expected value function (quatratic loss) we defined above. As a reminder, we are trying to reduce the error which in this case is the vertical distance from each point to the plane squared. In this way, we get a fit where there are many points which are close to the plane, and the points far away are reducted. Least squares linear regeresstion is the method which fits the data and moves the plane.

Since we have assumed a linear relationship between the response and the two predictors, the entire fitting problem reduces to estimating β_0, β_1,β_2, which we do using least squares linear regression. Least squares linear regression uses our expected value function to measure how well the model fits, penalizes error and finds the values of $\vec{\beta_{p+1}}$ which minimizes the error. We will learn more about least squares linear regression in the next chapter.

For now, let's work out the example numerically. This model is defined using three constants $(\beta_0, \beta_1, \beta_2)$. In the example, after we fit the data (we learn to to this in the next chapter), $\beta_0 = 2.911, \beta_1 = 0.188, \beta_2 = 0.0458$ so the overall equation would look like the following:

$$sales \approx 2.911 + 0.188 \times Radio + 0.0458 \times TV$$

In matrix form:

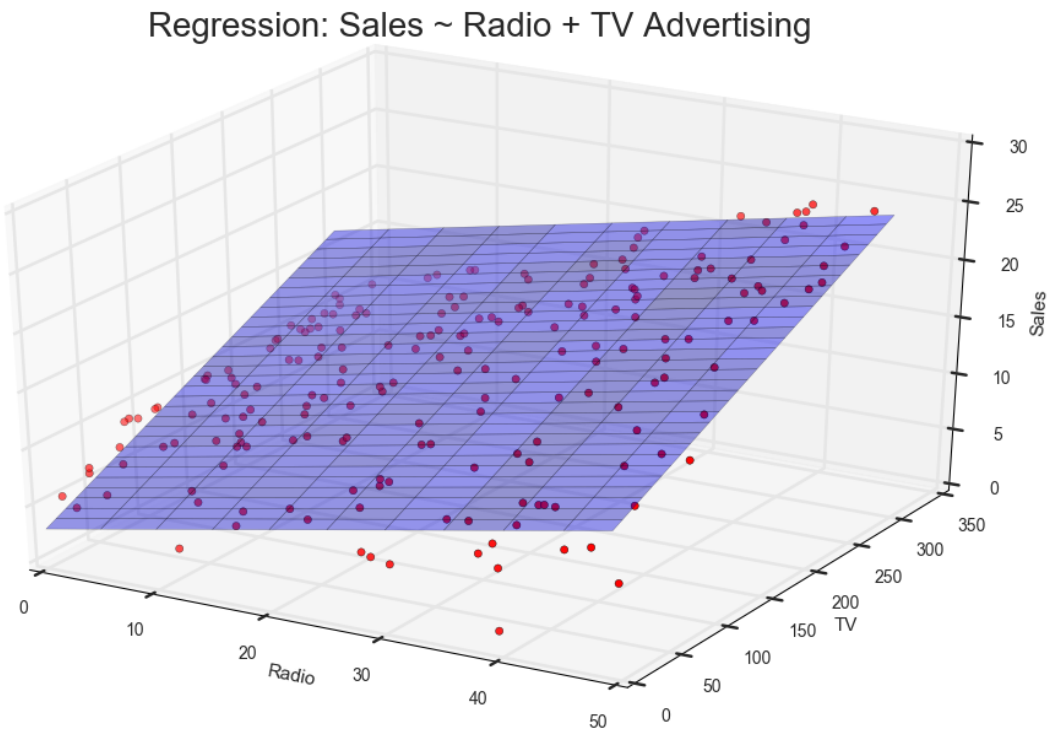

Figure 4:

$$\begin{bmatrix} Sales(\$MM) \\ 22.1 \\ 10.4 \\ \vdots \\ 13.4 \end{bmatrix} \approx 2.911 + 0.188 \begin{bmatrix} Radio(\$M) \\ 37.8 \\ 39.3 \\ \vdots \\ 8.6 \end{bmatrix} + 0.0458 \begin{bmatrix} TV(\$M) \\ 230.1 \\ 44.5 \\ \vdots \\ 232.1 \end{bmatrix}$$

Let's work out the first observation:

$$sales_1 \approx 2.911 + 0.188 \times Radio_1 + 0.0458 \times TV_1$$

$$\hat{Y_1} = sales_1 \approx 2.911 + 0.188 \times 37.8 + 0.0458 \times 230.1 \approx 20.61$$

If we compare that with $Y_1 = 22.1$, there is an absolute error of about $|Y - \hat{Y}| = 1.46$, and a quatratic loss of $E(Y - \hat{Y})^2 = 2.1439$ To get the total quadratic error (otherwise known as residual sum of squares or RSS), we would do this for all observations, and add it up. In equation form:

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

In this example, the $RSS = 556.91$ from fitting the plane. As you will notice there is some error in the prediction, and this is expected to happen. We can also potentially gain inference from fitting the linear model. We learn how to look at the statistical significance of linear regression in the next section, but for now we potentially have the knowledge that for every dollar spent on Radio advertisting, sales increase by about 188 dollars, and for every dollar spent on TV advertising, sales increase by about 46 dollars. We would

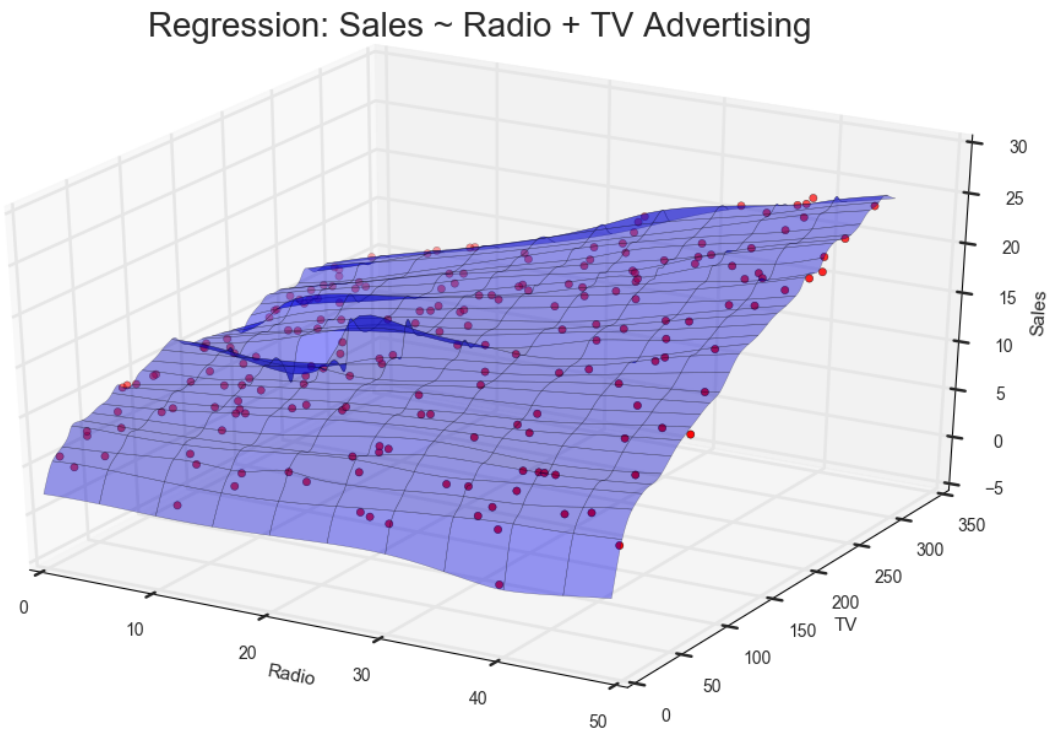

Figure 5:

want to validate the model, and verify that there is actually causation, but the initial correlation is a hint that this might be true!

The final step in modeling is evaluating the fit, and selecting the best model. Looking at the figure, the linear model is not quite right: the true f() has some curvature that is not captured in the linear fit. However, the linear fit still appears to do a reasonable job of capturing the positive relationship between Radio and TV advertising. Let's move onto non-parametric models and see if we can get a better fit.

Non-Parametric Methods

Non-parametric methods do not make explicit assumptions about the functional form of f() . Instead they seek an estimate of f () that gets as close to the data points as possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

The figure shows a thin-plate spline is used to estimate f. This approach does not impose any pre-specified model on f. It instead attempts to produce an estimate for f that is as close as possible to the observed data, subject to the fit. Now you might notice that is a small hump where it predicts a highs sales in the middle. This is an example of overfitting the data, which we discussed previously. It is an undesirable situation because the fit obtained will not yield accurate estimates of the response on new observations that were not part of the original training data set. For example, that hump might be attributed to high sales due to a

very good marketing campaign or a competitor's product tanking. Either way, it looks fishy and should be checked out. With hundreds of features, it makes it quite difficult to avoid over-fitting, so many times its better just to fit a simplier model. This brings us to the next subject: The trade off between prediction accuracy and model interpretability.