Introduction to Statistical Learning

CMSC498T: Introduction to Data Science, Spring 2015

The purpose of this lecture is to provide a brief overview of the statistical and machine learning techniques commonly used in data analysis. By the end of the term, you should be able to read papers that used these methods critically and analyze data using them. A common situation in applied sciences is that one has an independent variable or outcome Y and one or more dependent variables or covariates $X_1, \ldots, X_p$. One usually observes these variables for multiple "subjects".

Note: We use upper case to denote a random variable. To denote actual numbers we use lower case. One way to think about it: Y has not happened yet, and when it does, we see $Y = y$.

One may be interested in various things: What effects do the covariates have on the outcome? How well can we describe these effects? Can we predict the outcome using the covariates?, etc...

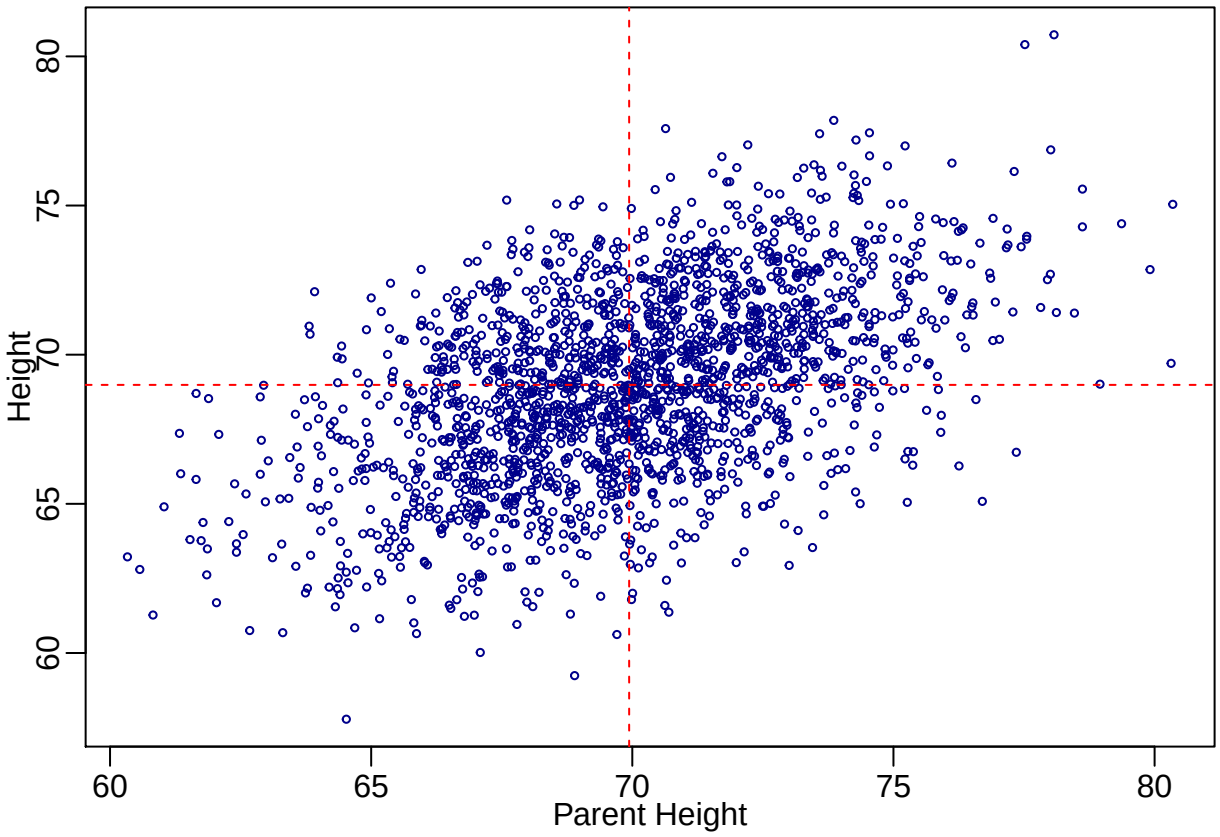
Linear Regression

Linear regression is the most common approach for describing the relation between predictors (or covariates) and outcome. Here we will see how regression relates to prediction.

Let's start with a simple example. Let's say we have a random sample of US males and we record their heights (Y) and the average height of their parents (X).

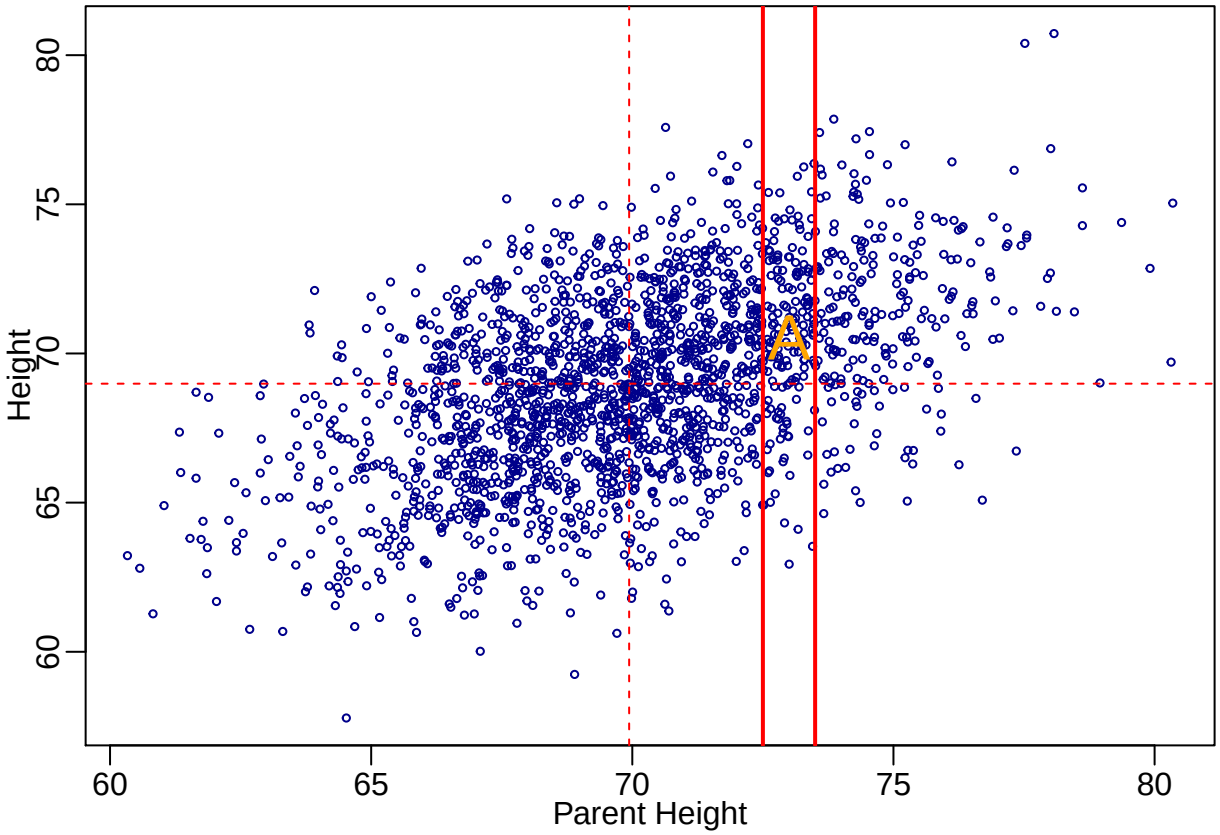
Say we pick a random subject. How would you predict their height?

What if I told you the average height of their parents? Would your strategy for predicting change?

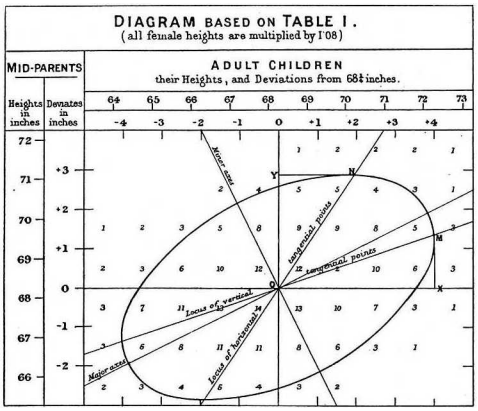


We can show mathematically that for a particular definition of "best", described below, the average is the best predictor of a value picked from that population. However, if we have information about a related variable, then the conditional average is best.

One can think of the conditional average as the average heights for all men with parents of particular average height.



In the case of height and parent height, the data actually look bivariate normal (football shaped) and one can show that the best predictor (the conditional average) of height given parent height is



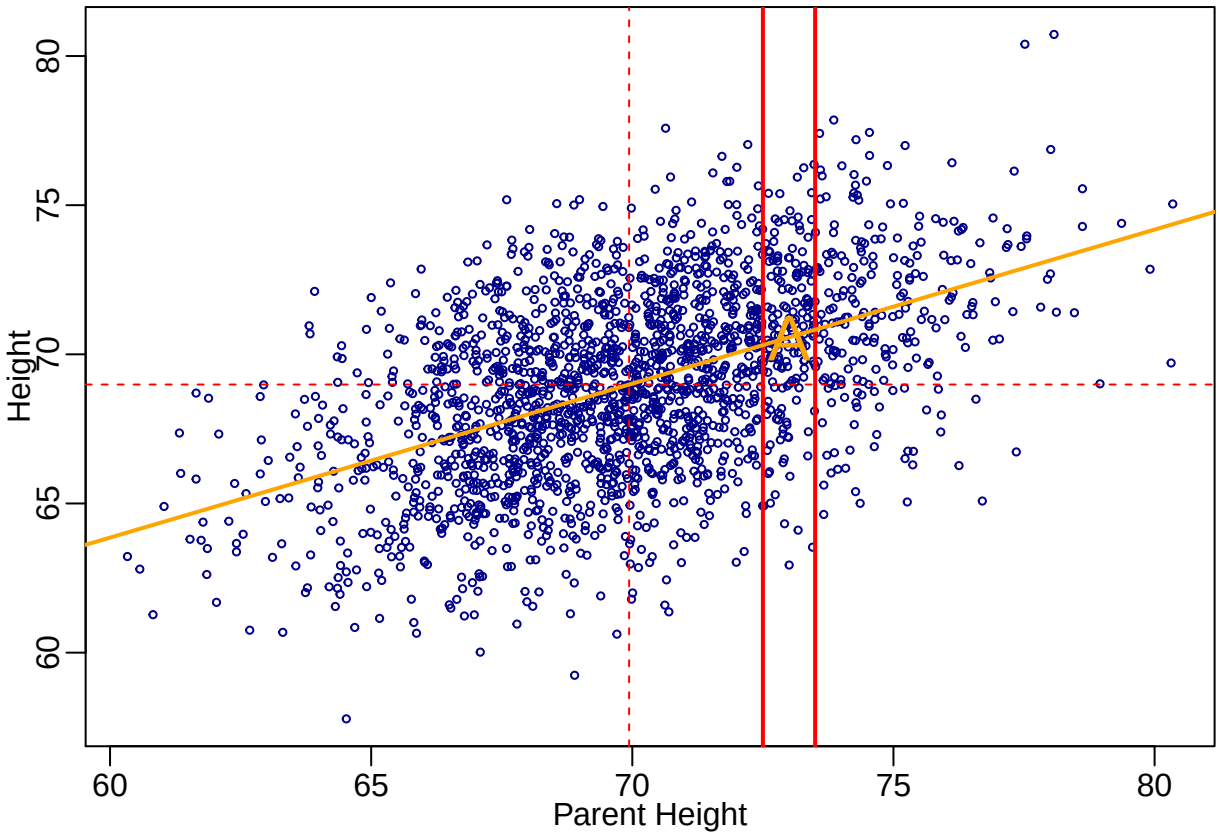
$$E[Y|X = x] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

with $\mu_X = E[X]$ (average parent height), $\mu_Y = E[Y]$ (average height), and where ρ is the correlation

coefficient of height and weight. Notice that this is a **linear function** of height!

If we obtain a random sample of the data, then each of the above parameters is substituted by the sample estimates and we get a familiar expression:

$$\hat{Y}(x) = \bar{Y} + r\frac{SD_Y}{SD_X}(x - \bar{X}).$$



Technical note: Because in practice it is useful to describe distributions of populations with continuous distributions we will start using the word *expectation* or the phrase *expected value* instead of average. We use the notation $E[\cdot]$. If you think of integrals as sums, then you can think of expectations as averages.

Notice that equation (1.1) can be written in this notation:

$$E[Y|X = x] = \beta_0 + \beta_1 x.$$

Because the conditional distribution of Y given X is normal, then we can write:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

with ϵ a mean 0, normally distributed random variable that is independent of X. This notation is popular in many fields because β_1 has a nice interpretation and its typical (least squares) estimate has nice properties.

When more than one predictor exists, it is quite common to extend this linear regression model to the multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

with ϵ as unbiased (0 mean) error independent of the X_j as before.

A drawback of these models is that they are quite restrictive. Linearity and additivity are two very strong assumptions. This may have practical consequences. For example, by assuming linearity one may never notice that a covariate has an effect that increases and then decreases. We will see various examples of this in class.

Linear regression is popular mainly because of the interpretability of the parameters. It allows us to perform *inference* about our measurements. E.g.,

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

However, the interpretation only makes sense if the model is an appropriate approximation of the natural data generating process. It is likely that the linear regression model from a randomly selected publication will do a terrible job at predicting results in data where the model was not trained on. Prediction is not really given much importance in many scientific fields, e.g. Epidemiology and Social Sciences. In other fields, e.g. Surveillance, Finance and web-commerce it is everything. Notice that in the fields where prediction is important, linear regression is not as popular.

Prediction

Methods for prediction can be divided into two general groups: continuous and discrete outcomes:

1. When the data is discrete we will refer to it as *classification*.
2. When the data is continuous we will refer to it as *regression*.

These seem very different but they have some in common. In this class, we will talk about the commonalities, but in general, we will discuss these two cases separately.

The main common characteristic in both cases is that we observe predictors $X_1, \ldots, X_p$ and we want to predict the outcome (or response) Y.

Note: I will use X to denote the vector of all predictors. So, X_i are the predictors for the i-th subject and can include age, gender, ethnicity, etc.

Note: Given a prediction method we will use $f(x)$ to the denote the prediction we get if the predictors are $X = x$.

Q: What are examples of prediction problems?

So, what does it mean to predict well? Let's look at the continuous data case first.

If I have a prediction $f(X)$ based on predictors X, how do I define a "good prediction" mathematically. A common way of defining closeness is with squared error:

$$L\{Y, f(X)\} = \{Y - f(X)\}^2.$$

We sometime call this the *loss function*.

Notice that because both Y and $f(X)$ are random variables, so is (2.2). Minimizing a random variable is meaningless because it is not a number. A common thing to do is minimize over the average loss, or the **expected prediction error**:

$$E_X E_{Y|X}[\{Y - f(X)\}^2|X].$$

For a given x, the expected loss is minimized by the conditional expectation:

$$f(x) = E[Y|X = x],$$

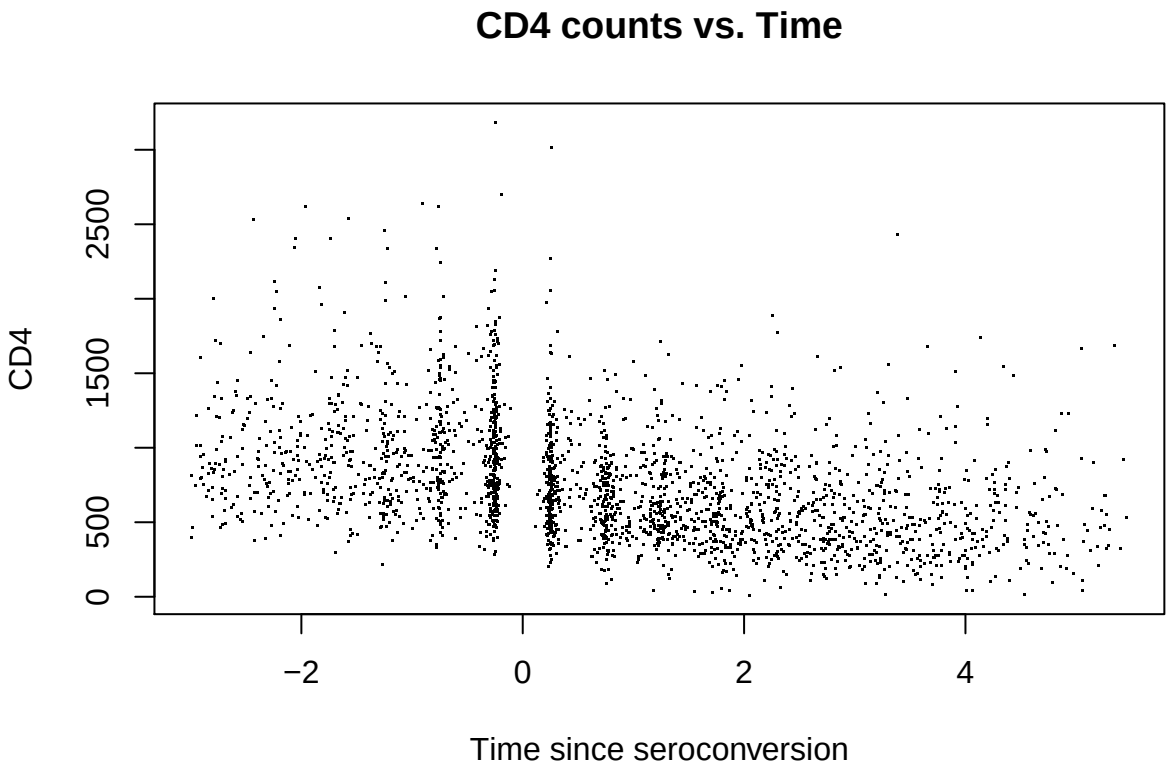
so it all comes down to getting a good estimate of $E[Y|X = x]$. We usually call $f(x)$ the *regression function.*

Note: For discrete problems we usually want a plausible prediction. Note $f(x)$ is typically a continuous number and not a class. We can take an extra step and define a prediction rule. For example, for binary outcomes, we can say: if $f(x) > 0.5$, I predict a 1, otherwise, predict 0. However, it is useful to change the loss function. More on this later.

Notice that if the regression model holds, then

$$f(X) = E[Y|X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

For Gaussian models, the solution is the same for least squares. However, many times, it is hard to believe that the linear regresion model holds. A simple example comes from AIDS research:



Other settings

A major focus of this class is prediction, or *supervised learning.* However, we will also see a few other learning settings. For instance, suppose we only observe vectors of random variables, $X_1, \ldots, X_p$ but no outcome Y? In this case we still want to find some informative structure (e.g. *clustering*). This setting is called *unsupervised learning.* We can include probability density estimation under this setting.

Terminology and notation

We will be mixing the terminology of statistics and computer science. For example, we will sometimes call Y and X the outcome/predictors, sometimes observed/covariates, and even input/output.

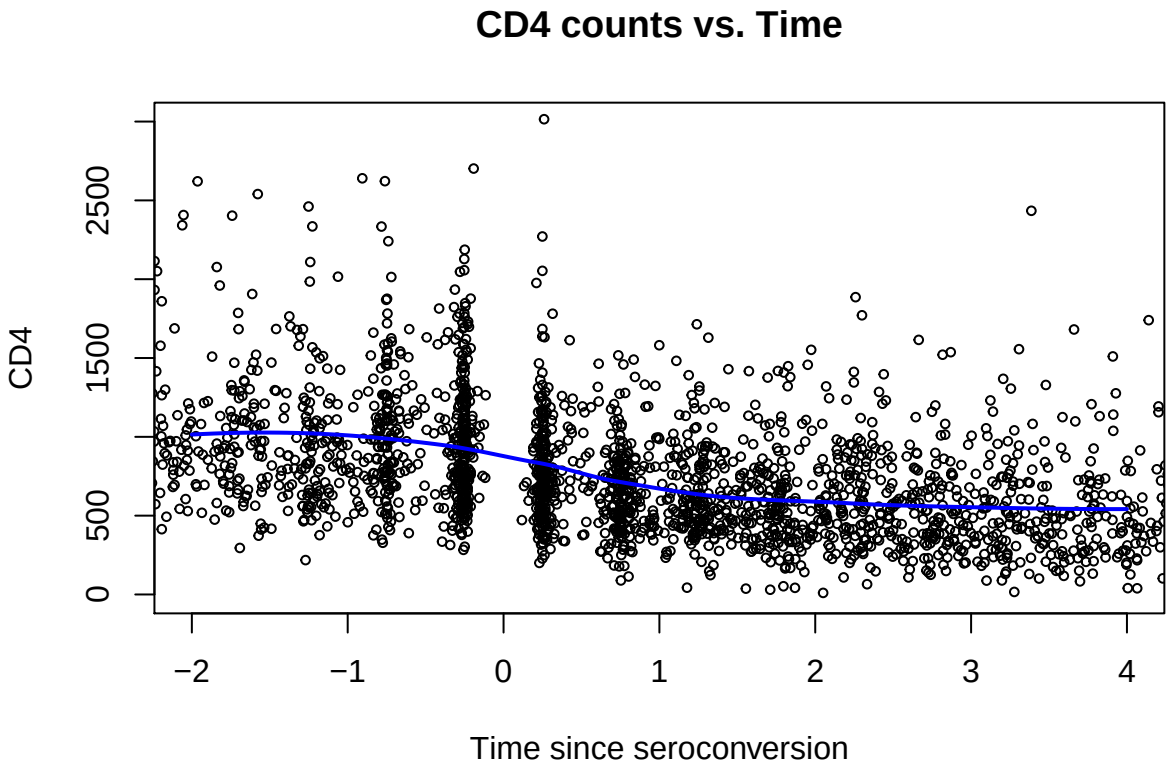
We will sometimes denote predictors with X and outcomes/responses with Y (quantitative) and G (qualitative). Notice G are not numbers, so we cannot add or multiply them.

Height and weight are *quantitative measurements*. These are sometimes called continuous measurements.

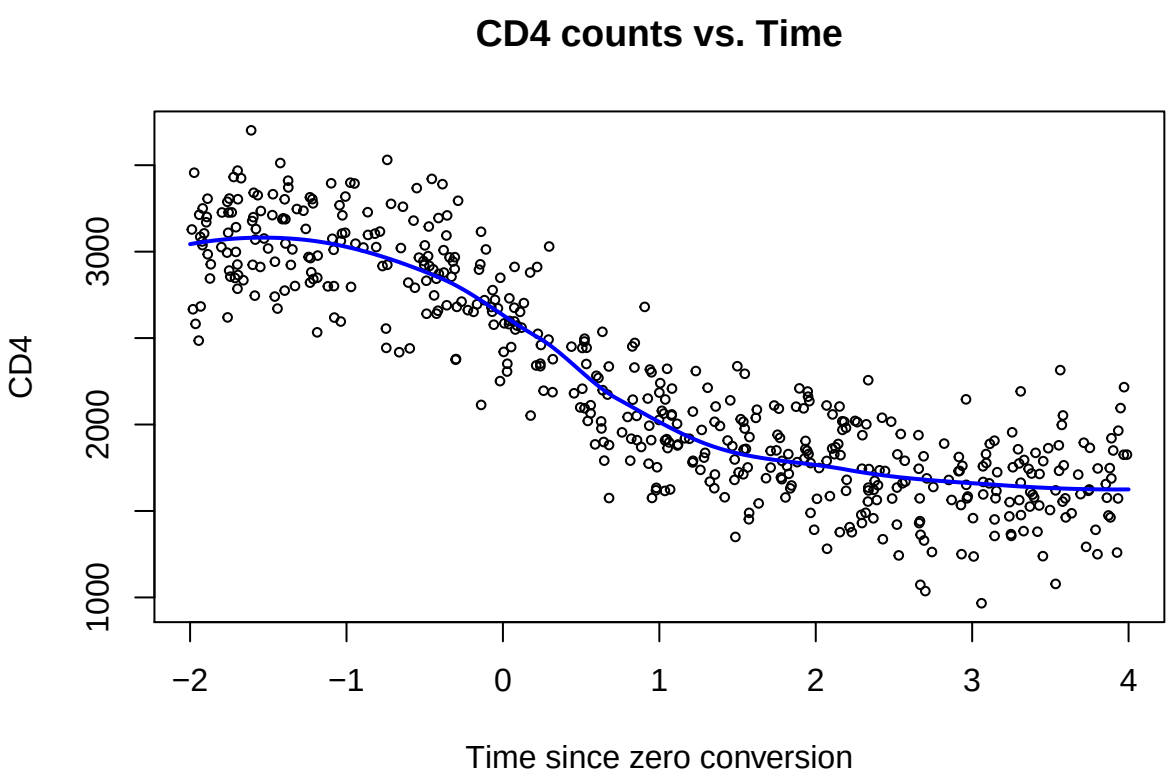
Gender is a *qualitative measurement*. They are also called categorical or discrete. This is a particularly simple example because there are only two values. With two values we sometimes call it *binary*. We will use G to denote the set of possible values. For gender it would be $G = \{Male, Female\}$. A special case of qualitative variables are *ordered qualitative* where one can impose an order. With men/women this can't be done, but with, say, $G = \{low, medium, high\}$ it can.

A regression problem

Recall the example data from AIDS research mentioned previously. Here we are plotting the data along with a curve from which data could have plausibly been generated.



For now, let's consider this curve as truth and simulate CD4 counts from it. We will use this simulated data to compare two simple but commonly used methods to predict Y (CD4 counts) from X (Time), and discuss some of the issues that will arise throughout this course. In particular, what is overfitting, and what is the bias-variance tradeoff.



Linear regression

Probably the most used method in data analysis. In this case, we predict the output Y via the model

$$Y = \beta_0 + \beta_1 X.$$

However, we do not know what β_0 or β_1 are.

We use the training data to *estimate* them. We can also say we train the model on the data to get numeric coefficients. We will use the hat to denote the estimates: $\hat{\beta}_0$ and $\hat{\beta}_1$.

We will start using β to denote the vector $(\beta_0, \beta_1)'$. We would call these the *parameters* of the model.

The most common way to estimates βs is by least squares. In this case, we choose the β that minimizes

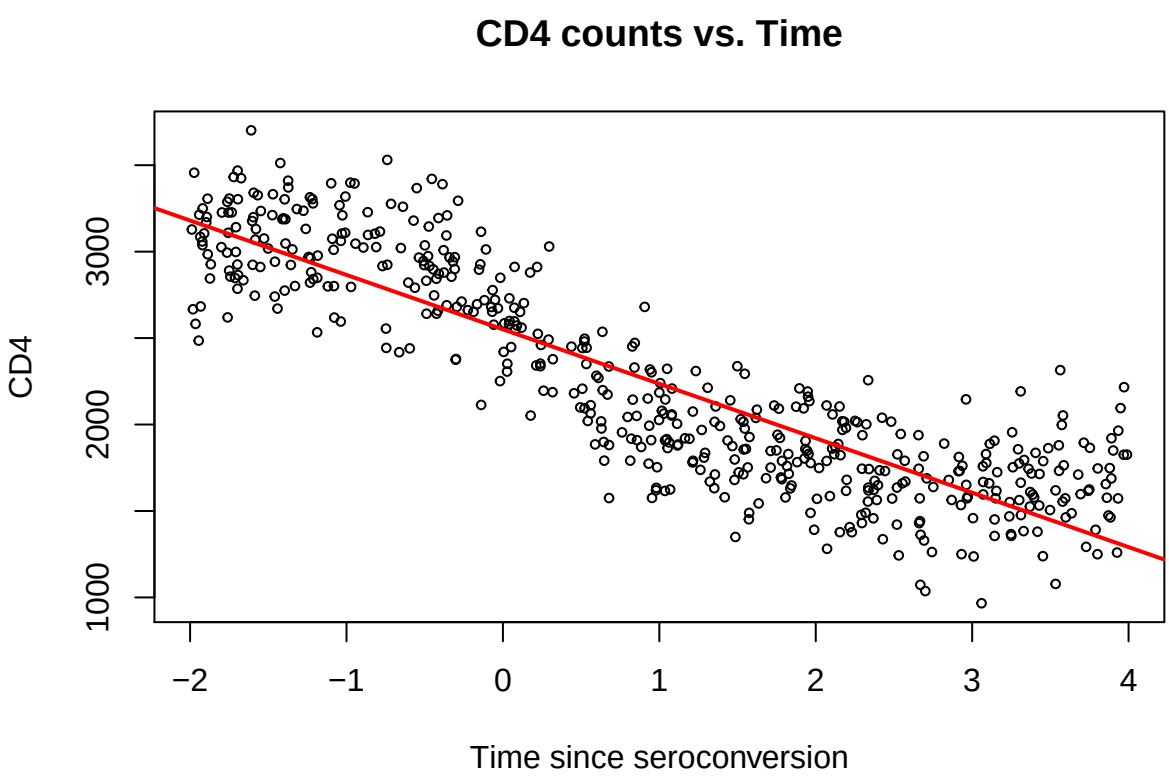
$$RSS(\beta) = \sum_{i=1}^{N} \{y_i - (\beta_0 + \beta_1 X_i)\}^2.$$

If you know linear algebra and calculus you can show that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} X'y$.

Notice we can predict Y for any X:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The next Figure shows the prediction graphically. However, the data seems to suggest we could do better by considering more flexible models.



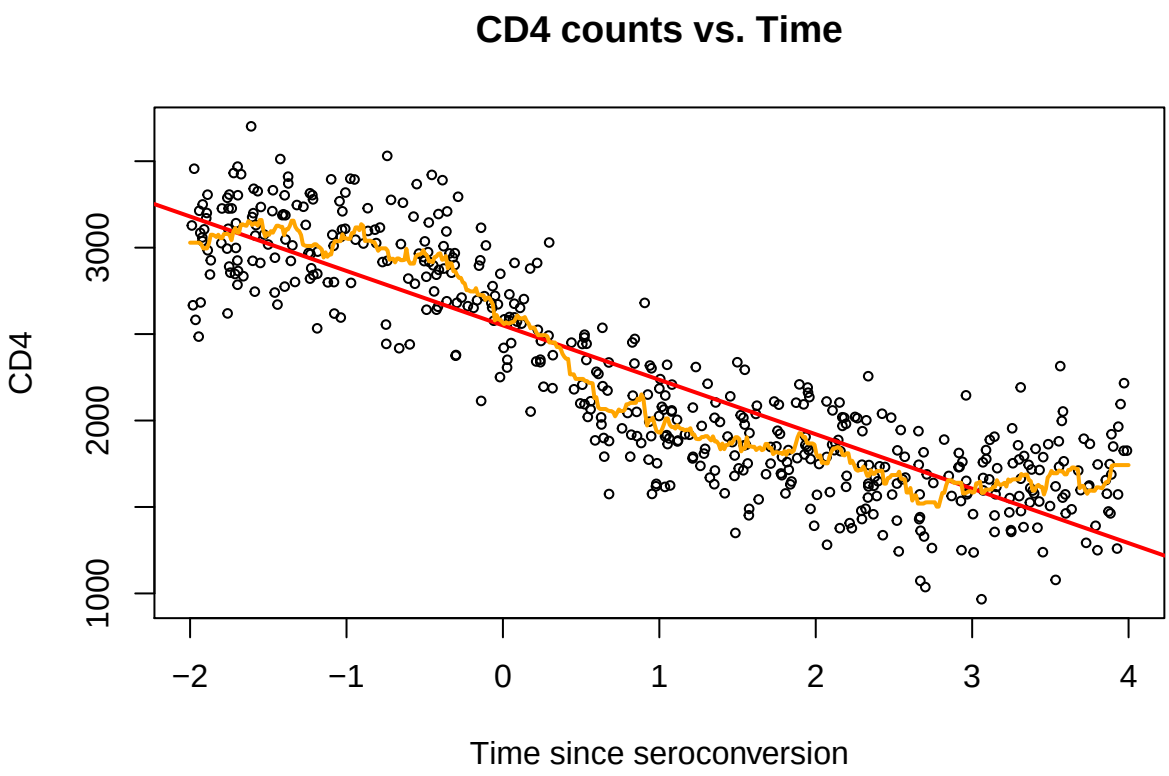
K-nearest neighbor

Nearest neighbor methods use the points closest in predictor space to x to obtain an estimate of Y. For the K-nearest neighbor method (KNN) we define

$$\hat{Y} = \frac{1}{k} \sum_{x_k \in N_k(x)} y_k.$$

Here $N_k(x)$ contains the k-nearest points to x. Notice, as for linear regression, we can predict Y for any X.

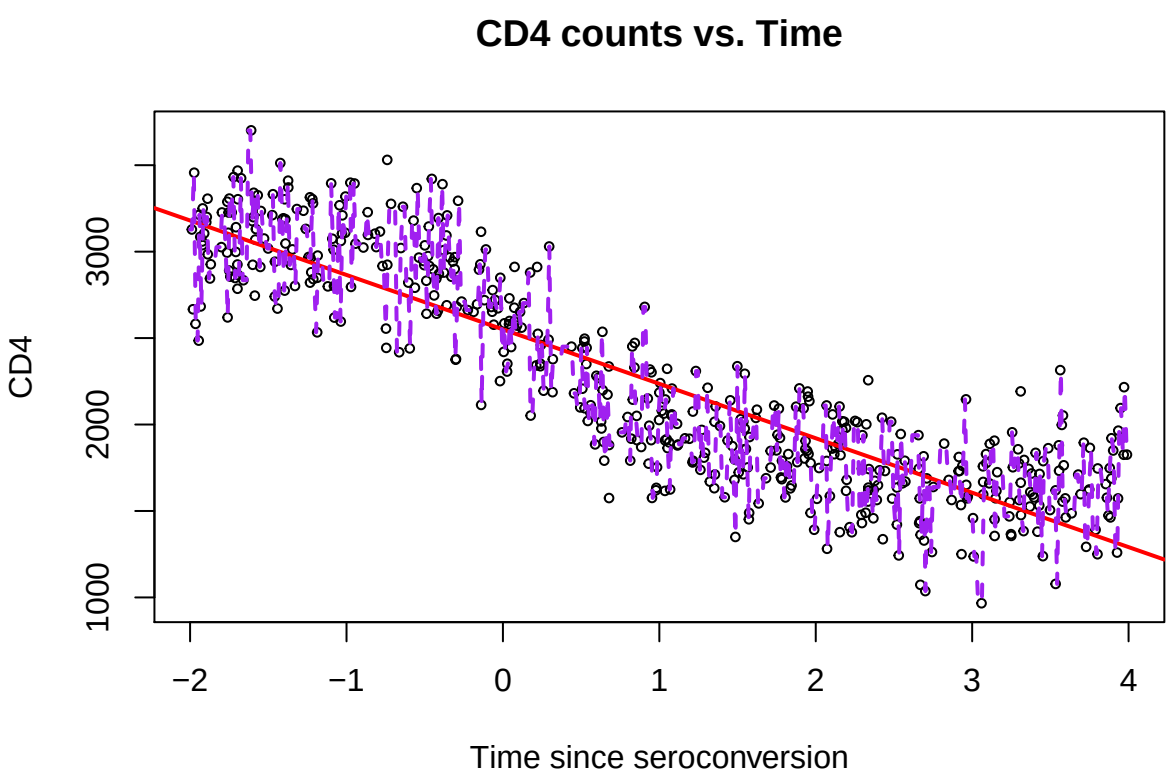
In the next Figure we see the results of KNN using the 15 nearest neighbors. This estimate looks better than the linear model.



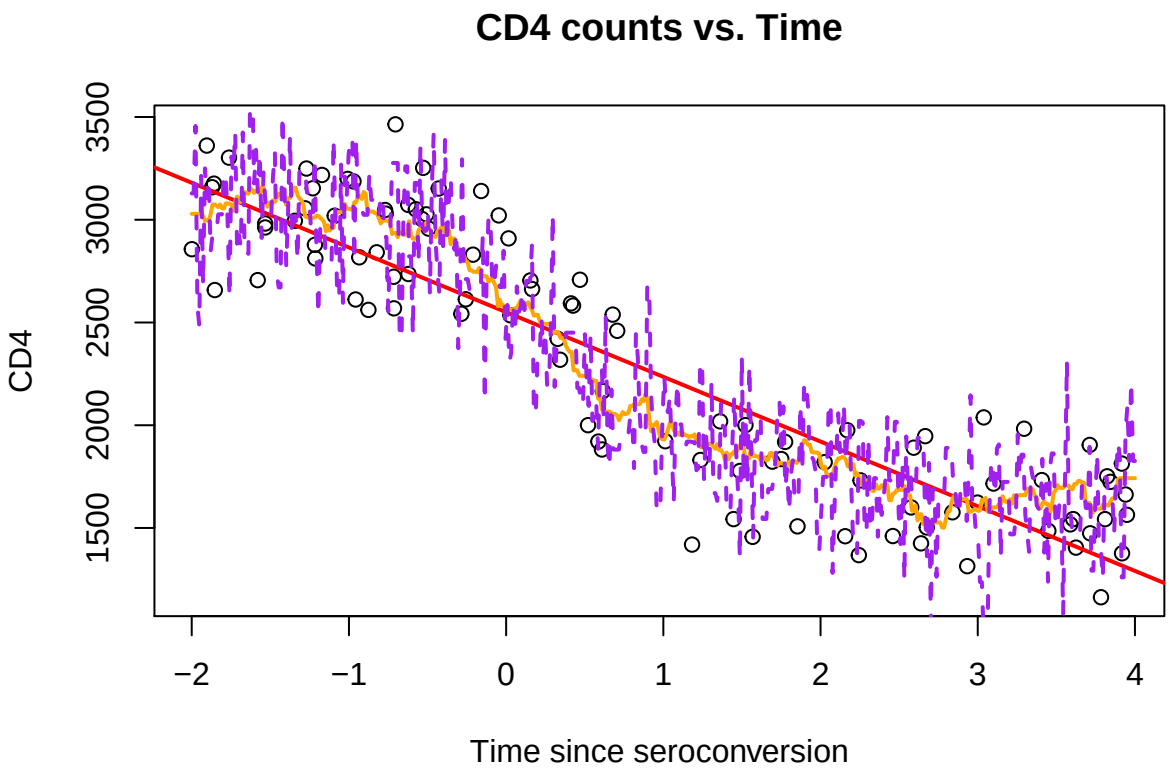
We do better with KNN than with linear regression. However, we have to be careful about *overfitting*.

Roughly speaking, *overfitting* is when you mold an algorithm to work very well (sometimes perfect) on a particular data set forgetting that it is the outcome of a random process and our trained algorithm may not do as well in other instances.

Next, we see what happens when we use KNN with k=1. In this case we make no mistakes in prediction, but do we really believe we can do well in general with this estimate?



It turns out we have been hiding a *test* data set. Now we can see which of these trained algorithms performs best on an independent test set generated by the same stochastic process.



We can see how good our predictions are using RSS again.

Method	Train set	Test set
Linear	99.70	93.58
K=15	67.41	75.32
K=1	0.00	149.10

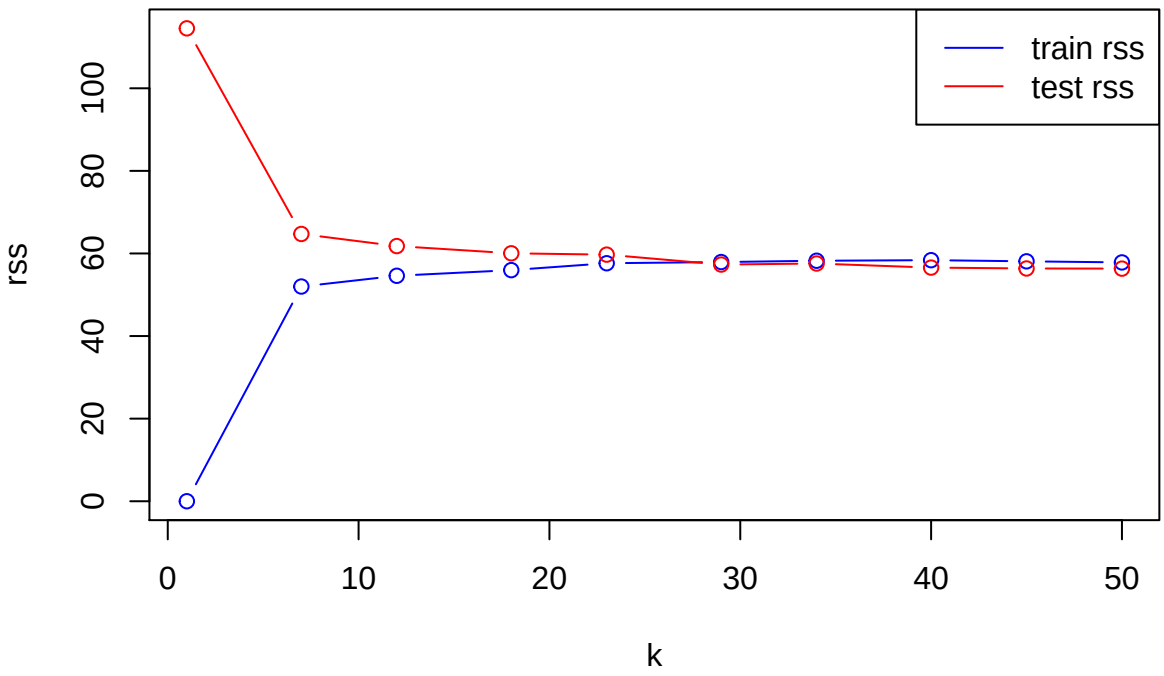
Notice RSS is worse in test set than in the training set for the KNN methods. Especially for KNN=1. The spikes we had in the estimate to predict the training data perfectly no longer helps.

So, how do we choose k? We will study various ways. First, let's talk about the bias/variance trade-off.

Smaller k give more flexible estimates, but too much flexibility can result in over-fitting and thus estimates with more variance. Larger k will give more stable estimates but may not be flexible enough. Not being flexible is related to being biased.

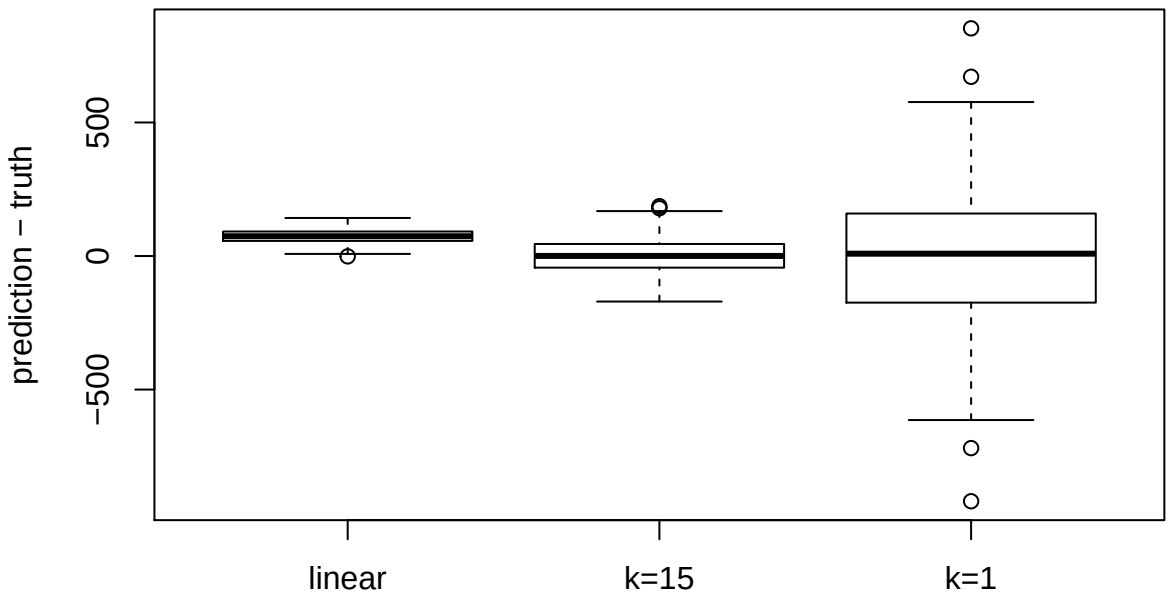
The next figure shows the RSS in the test and training sets for KNN with varying k. Notice that for small k we are clearly overfitting.

```
## 171218232934404550
```



An illustration of the bias-variance tradeoff

The next figure illustrates the bias/variance tradeoff. Here we make boxplots of $f(1) - \hat{f}(1)$, where $\hat{f}(1)$ is the estimate for each method trained on 1000 simulations.



We can see that the prediction from the linear model is consistently inaccurate. That is, it is biased, but stable (little variance). For $k = 1$ we get the opposite, there is a lot of variability, but once in a while it is very accurate (unbiased). For $k = 15$ we get a decent tradeoff of the two.

In this case we have simulated data as follows: for a given x

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

where $f(x)$ is the "true" curve we are using and ϵ is normal with mean zero and some variance σ^2.

We have been measuring how good our predictions are by using RSS. Recall that we sometimes refer to this as a *loss function*. Recall also that for this loss function, if we want to minimize the **expected prediction error** for a given x:

$$E_{Y|X=x}[\{Y - f(X)\}^2|X = x],$$

we get the conditional expectation $f(x) = E[Y|X = x]$. With some algebra we see that the RSS for this optimal selection is σ^2 in our setting. That is, we can't do better than this, on average, with any other predictor. This is called *irreducible error* in the book.

Notice that KNN is an intuitive estimator of this optimal predictor. We do not know the function $E[Y|X = x]$ looks like so we estimate it with the y's of nearby x's. The larger k is, the less precise my estimate might be since the radius of x's I use for is larger.

Predictions are not always perfect.