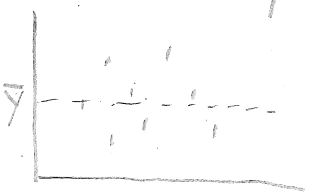
Why do we do data science? predict the future

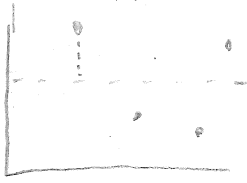
How do we do data science? Take some sample observations, fit a model to them, and then use the model to make predictions. Typically there is a target variable, which we will also call the dependent variable, that we don't know & want to predict, and some other information about the same object that we do know. We need the model to relate what we do know & predict the value of target variable.

What model(s) did we fit to data in 131? 1-D case Point or horizontal line (mean) & a line (OLS). 2-D case

We will use illustrations of 2-D data sets, but the principles apply to data of any number of dimensions.

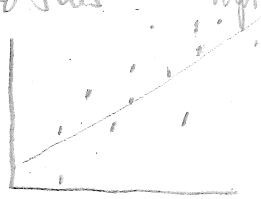


Modeling the data using the mean. No matter what x is, we predict $\bar{y}$. x is superfluous. What is the mean? Measure of center? What is standard deviation? Measure of variability around the mean. We can also call this the RMSE in the 1-D case. What is the formula (you have to know this now)? $\sqrt{\frac{1}{N-1}\sum_i (y_i-\bar{y})^2}$ Why the N-1? What is I called in this formula? Python calls it ddof, short for delta degrees of freedom. In statistics, degrees of freedom is the number of independent data points - the number of parameters that can vary in creating the model to fit them. In this case, only the mean can vary (different sample, different mean). We are really talking about the estimated mean, because unless we have the entire population, our calculated mean is an estimate of the real mean. Look again:



Some one come up here & draw a $y_i-\bar{y}$. What is this called? A residual. Sometimes the error, but this is incorrect. In statistics, the error is the unknowable true value. Confusingly though, the mean of the sum of the squares of these things is called the SSE or sum of squared error. Just have to get used to it. If you want fit error in there, call the residual the "predicted error." That is correct.

What other model did we fit to 2-D data in 131? Lines that were not constrained to be horizontal. What did we call the mathematical procedure to this? Simple linear regression or ordinary least squares.



Can anybody explain to me how this procedure works? It minimizes the SSE for the fitted line. This is what we define as the 'best fit line'. There are other possible choices - we could minimize the perpendicular (drawn) squares instead of the vertical squares, but the math gets to be much more of a pain & OLS gives us really nice fits, so why bother.

Let's step back for a second. What is a statistic? A number that describes some data. When someone calculates & reports a statistic, they often present the statistic & another number. What's the other number for? Tell us how much significance we should attach to the statistic. Often a p-value is reported - anyone know the definition? The probability that a result this extreme or more would result if the null hypothesis were true. What's the null hypothesis? In general terms, "there's nothing to see here, move along." E.g. our new drug doesn't really work, the climate isn't really changing, our model doesn't predict who will default any better than a coin flip, etc. Let's add a practical definition for a model: one or more equations that describes some data: that's what it always comes down to. We use these equations as the components of a procedure that we use to predict the value of a target variable for an object/instance given some other data about the object/instance. To summarize: we make some observations (get data) about some objects/instances (in the programming sense - they don't have to be physical objects, just objects with shared attributes/characteristics) that include the values of the target variable, then we make a model that predicts the value of the target variable for new instances given the rest of the attributes of the new instance. Last semester, the model we learned was the line. We used OLS to fit a line to data, so now we are going to pick up from there.

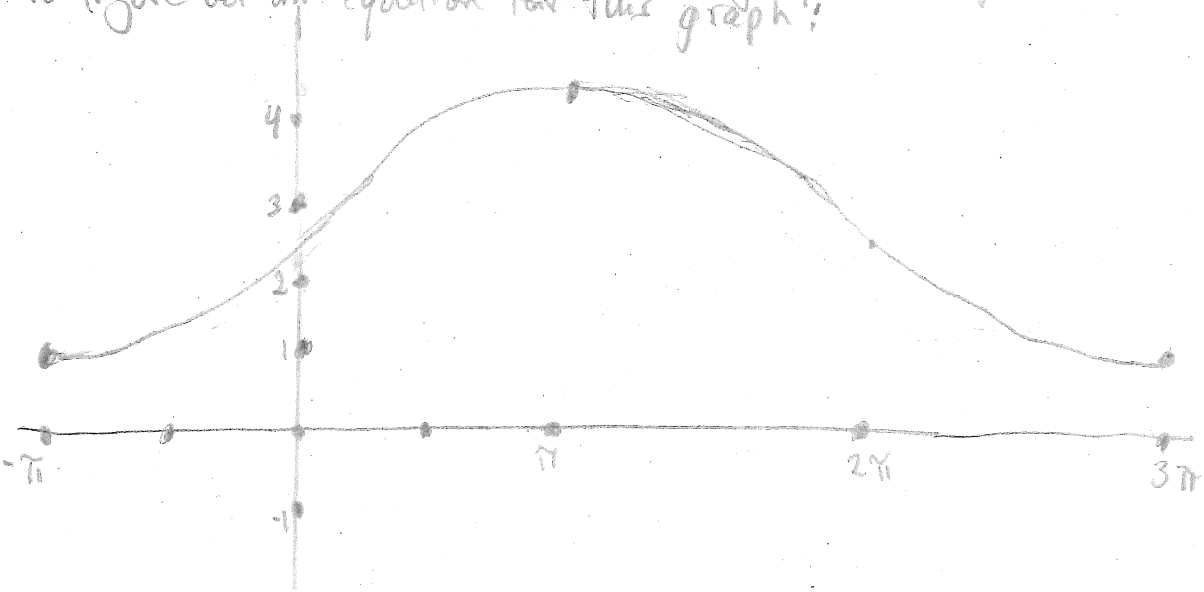
What do you think of modeling this data w/ the mean? A line? What do you suggest? A parabola. Ok, an aside. It is normal in our society to be afraid of math because it is so poorly taught. If that describes you, you have to let that go right now. You're a data scientist - math is exciting. Every new math thing is something you want to absorb, to make your own. You want to own that math. I'm not saying that math is easy; I am saying you can't be afraid of it. Ok, so here we go! We used statsmodels ols routine to fit lines to collections of points. In our case, we regressed our target variable, which I think in one case was Arctic sea ice extent, with the independent variables being time. But OLS works fine for more than one independent variable, so suppose we had a measure of storminess, we could have regressed ice extent vs. time & storminess to get an equation like:

$$Extent = a(time) + b(storminess) + c$$

This is still called linear regression because all of the parameters that we are varying to fit the model (what are they? a, b, c) are linear. There is no a^2 or b^5 or $\sin c$ or anything like that. So OLS still works fine. And we can use a trick to take advantage of this. Suppose we want to fit a parabola to the above points. The parabola will have the form of a quadratic $y = ax^2 + bx + c$. So we can do part of the math ourselves, namely squaring x, and treating these x^2 values as a separate value, and let statsmodels do its thing. (Do the example). Now we have our model, i.e. best-fit values for a, b, c. Now we need some kind of a measure of how good this model is that people can use to decide whether it's useful or not. Any ideas? For fitted curves, we often use measures of goodness of fit. In other words, a number that tells us how close our model is to the data points. What is our 1-D measure of goodness of fit? Std dev. What about 2-D? Pearson's r, RMSE, R^2 coefficient of determination - in the linear case = the % of variance explained by the independent variable. In this case $R^2 = r^2$ & you will sometimes see it written that way. Statsmodels calculates MSE (what do we do if we

want RMSE? Take its root) and R^2 for us. It also calculates a statistic, the F-statistic, that describes our model. The documentation is opaque, but I believe this number (calculated using the F-test) compares the fit of our model to the fit of the mean of y-values. In any case, the bigger the number, the better. As with many statistics, statsmodels also reports a p-value, called f-pvalue, that tells us how much weight we should give to the F-statistic. (Show all this stuff)

Sometimes we need to fit an equation with nonlinear parameters, for example $y = a\sin(bx+c)+d$. b & c are inside an argument to sin, so we can't use our trick. In this case, we must move beyond OLS, (O. the curve_fit eg. with default params). Ok, that's a terrible fit. So what happened? Fitting these curves are examples of optimization problems. Specifically, minimization problems. We are minimizing so measure of the distance between our curve & the points. In this case, we can't directly calculate the optimal values of our parameters like we OLS does. So we have to start with estimated values and vary our params until we get a best fit. But if our initial estimate (in this case, all 1's) is way off, then we'll hit a local minimum before we get to the best values. So we need a better starting point. At this point, we need to learn some principles of graphs. Suppose we want to figure out an equation for this graph!



Well, first we have to know enough math to recognize that we can we do this cos or sin and we need to know how to graph the starting equation $y = \sin x$ (or $\cos x$). <Add the starting equation>. We are going to fit an equation of the form $y = a\sin(bx+c)+d$ to this curve. Any suggestions as to what a should be? Well what does a do? Whether or not someone knows, show the example on a separate graph of $y=\sin x$, $y=2\sin x$, $y=3\sin x$. So a stretches/compresses the curve vertically, starting range = 2, goal range = 3, so $a = 1.5$. What does d do? Do an example. So it moves the vertical center of the curve up or down. So what was 0 is now 2.5, so $d = 2.5$. Now this part gets tricky. It would be different if we wrote our starting form as $y = a\sin(b(x+c))+d$. It's a different way of writing the same thing, but it changes the order of operations, and with it, how we need to think about what we're doing it. c moves the curve horizontally. With our formulation, we need to think of this happening prior to b, which stretches/compresses the curve horizontally. In the starting equation the inflection pt with positive second derivative is at $x=0$. It still is, so $c=0$. One cycle starts as 2π, but is 4π in the new curve, so $b = \frac{1}{2}$. We have $y = 1.5\sin\left(\frac{x}{2}\right)+2.5$