Lecture Notes part C

PSY V0500, Statistical Methods in Psychology

Kevin R Foster, the Colin Powell School at the City College of New York, CUNY

Spring 2024

Table of Contents

Simple Machine Learning

From basic notions of mean and standard deviation, we can quickly move to some simple types of machine learning. This is a great example of a very simple idea that has some fancy-sounding terminology. The idea is that if you want to classify a new observation then the easiest guess is to ask how old observations that were very near were classified. "Birds of a feather flock together," or "You're judged by the company you keep."

There are many possibilities, where we gather data on some preliminary values and try to predict something else. If we have a big dataset on past students who were admitted or not to a certain program, we could use this data to predict future admits. Lots of marketing might use this sort of algorithm: if consumers are similar by some characteristics then they might be similarly receptive to a certain ad.

The machine learning technique called "K Nearest Neighbors" or "k-nn" uses other observations that are "nearby" to try to classify new observations.

What does "near" mean? If we have a list of numeric data then the temptation is to just use simple distance (typically Euclidean). There are two aspects to this choice: first, what variables are helpful in classification; second, how are these variables scaled? The choice of what variables is a bit tricky since we want to find some good ones but not too many (too many, relative to the size of the data, gets into the "curse of dimensionality" and there are usually few neighbors). That usually requires a bit of background knowledge – this is called "machine learning" but it's actually strongly human-controlled machine learning (so cyborg learning?).

The second part of "near" is a bit more subtle: the scaling of each variable is important. If a college is classifying high school students as either admit or not, they might use GPA and SAT. If HS GPA is on a scale of 0-4 then for selective colleges most of the relevant admissions will have GPA from 3.5-4. SAT scores on the other hand (for now assume they're math plus verbal) could have differences of hundreds of points. So the SAT score variation will swamp the GPA variation. (This is why some people think of scores on standardized tests as their percentile.)

Detour on Ranking

We often see statistics reported that rank a number of different units based on a number of different measures. For instance, these could be the US News ranking of colleges, or magazine rankings of city livability, or sports rankings of college teams, or any of a multitude of different things. We would hope that statistics could provide some simple formulas; we would hope in vain.

Education: College rankings try to combine student/faculty ratios, measures of selectivity, SAT scores, GPA; some add in numbers of bars near campus or the prestige of journals in which faculty publish. What is best? School teachers face efforts to rank them, by student test score improvements as well as other factors; schools and districts are ranked by a variety of measures.
Sports might seem to have it relatively easy since there is a single ranking given by pre-arranged rules, but still fans can argue: a team has a good offense because they scored a lot (even though some other team won more games); some players are better on defense but worse on offense. Sports Illustrated tried to rank the 100 all-time best sports stars, somehow comparing baseball player Babe Ruth with the race horse Secretariat! Most magazines know that rankings drive sales and give buzz.
Food nutrition trades off calories, fat content, fiber, vitamin and mineral content; who is to say whether kale or blueberries are healthier? Aren't interaction effects important? Someone trying to lose weight would make a very different ranking than someone training for a marathon.

Sustainability or "green" rankings are difficult: there are so many trade-offs! If we care about global warming then we look at CO_2 emissions, but what about other pollutants? Is nuclear power better than natural gas? Ethical consumption might also consider the material conditions of workers (fair-trade coffee or no-sweatshop clothing) or other considerations.
Politics: which political party is better for the economy? Could measure stock returns or unemployment rate or GDP growth or hundreds of others. Average wage or median earnings (household or individual)? Each set of measures could give different results. You can try this yourself, get some data from FRED (http://research.stlouisfed.org/fred2/) and go wild.

In the simplest case, if there is just a single measured variable, we can rank units based on this single measure, however even in this case there is rarely a clear way of specifying which rankings are based on differences that are large and which are small. (The statistical theory is based on "order statistics.") If the outcome measure has, for example, a normal distribution, then there will be a large number of units with outcomes right around the middle, so even small measurement errors can make a big difference to ranking.

In the more complicated (and more common) case, we have a variety of measures of outcomes and want to rank units based on some amalgamation of these outcomes. Economic theory has a very strong result here: a bit of math can prove that there is no way to generate a function for a group of people that completely and successfully takes account of the information of individual choices. (This result is due to CCNY alumnus and Nobel Laureate Ken Arrow.) It can't be done. In general many rankings can be substantially changed by adding factors or even changing the units of certain of the factors (changing the measure of "near" as discussed before).

Many rankings take an equal weighting of each item, but there is absolutely no good reason to do this generally: why would we believe that each measure is equally valid? Some rankings might arbitrarily choose weights or take a separate survey to find weights (equally problematic!). You could average what fraction of measures achieve some hurdle. But there's no reason to think that's better.

One possible way around this problem might seem to be: just ask for people's rankings (let them figure out what weights to use in their own utility functions) and report some aggregation. However here again there is no single method that is guaranteed to give correct aggregations (this is the Ken Arrow result again). Some surveys ask people to rank units from 1-20, then add the rankings and the unit with the lowest number wins. But what if some people rank number 1 as far ahead of all of their competitors, while others see the top 3 as tight together? This distance information is omitted from the rankings. Some surveys might, instead, give 10 points for a #1 ranking, 8 points for #2, and so on – but again this presupposes some distance between the ranks.

This is not to say that ranking is hopeless or never informative, just that there is no single path that will inerrantly give the correct result. Working through various rankings, an analyst might determine that a broad swathe of weights upon the various measures would all give similar rankings to certain outliers. It would be useful to know that a particular unit is almost always ranked near the top while some other one is nearly always at the bottom.

Cathy O'Neil's book, *Weapons of Math Destruction*, gives many more examples of problems that arise.

Other Ignorant Beliefs

While I'm working to extirpate popular heresies, let me address another one, which is particularly common when the Olympics roll around: the extraordinary belief that outliers can give useful information about the average value. We hear these judgments all of the time: some country wins an unusual number of Olympic medals, thus the entire population of the country must be unusually skilled at this task. Or some gender/race/ethnicity is overrepresented in a certain profession

> thus that gender/race/ethnicity is more skilled on average. Or a school has a large number of winners of national competitions, thus the average is higher. Really?
>
> Statistically speaking, the extreme values of a distribution depend on many parameters such as the higher moments. If I have two distributions with the exact same mean, standard deviation, and skewness, but different values of kurtosis, then one distribution will systematically have higher extremes (by definition of kurtosis). So in general it is not true to infer that a higher number of extreme values implies a higher mean. But people do.

Rankings can be shifted by different values of "near" as can machine learning algorithms. It is up to you to learn about how to use these most adroitly.

The variation in a measure is sometimes called its "information". Consider even a simple case where students' grades in a class are determined by even weighting of 2 exams. If scores on one exam are much more variable than scores on the other exam then they don't actually end up contributing equal weight to student ranking. (Think of the limiting case where everyone gets the same score on one exam, therefore it has no contribution to ranking even if it is given 50% weight.)

A common way to manage this is to standardize the predictors (subtract mean and divide by standard deviation) or scale them to be all in the [0,1] interval, although this is far from perfect. There is an art to choosing predictors. Although it might not seem obvious, this is essentially the same problem as with rankings.

I provided a Lab with some detail of code for k-nn.

It uses a technique that we'll often return to: splitting the data into a training set and a test set. If the point of a model is to predict some data, then I want to test it out on some data that was not used for training. For example you've doubtless taken classes with various types of exams. Sometimes the instructor will give students a number of practice problems then the exam would consist of some of those problems. Other times the instructor will give practice problems but then the exam is new problems that are related to the practice but not identical. I think you'd agree that the second type is more difficult!

We want to test our models similarly and don't just reuse data to give an easy test. We take out some of the data and don't use that in the estimation. The data used for estimation is the "training" data, that we use to train the model. The test data is separate, used to test how well that model performs on data that it hasn't seen before. Here we use 80% of the data as the training set and the remaining 20% as the test set.

The "set seed" command is a bit of magic that lets us take a random sample but if you do it again the computer would take the same "random" sample. The computer doesn't actually take a random sample but it is actually pseudo-random where complicated algorithms create numbers that look random in many ways but are actually deterministic so if we start from the same value then we get the same list of random numbers. The "seed" sets that starting point. You might think, why not just take the first 80% of the sample, but that would depend on the assumption that the ordering of data is random. Many datasets have structure so the observations might be ordered in some way.

The program finishes by using the knn routine. It can use different numbers of nearest neighbors so experiments with using 1, 3, 5, 7 or 9 nearest neighbors for the classification and reports how accurate each one is.

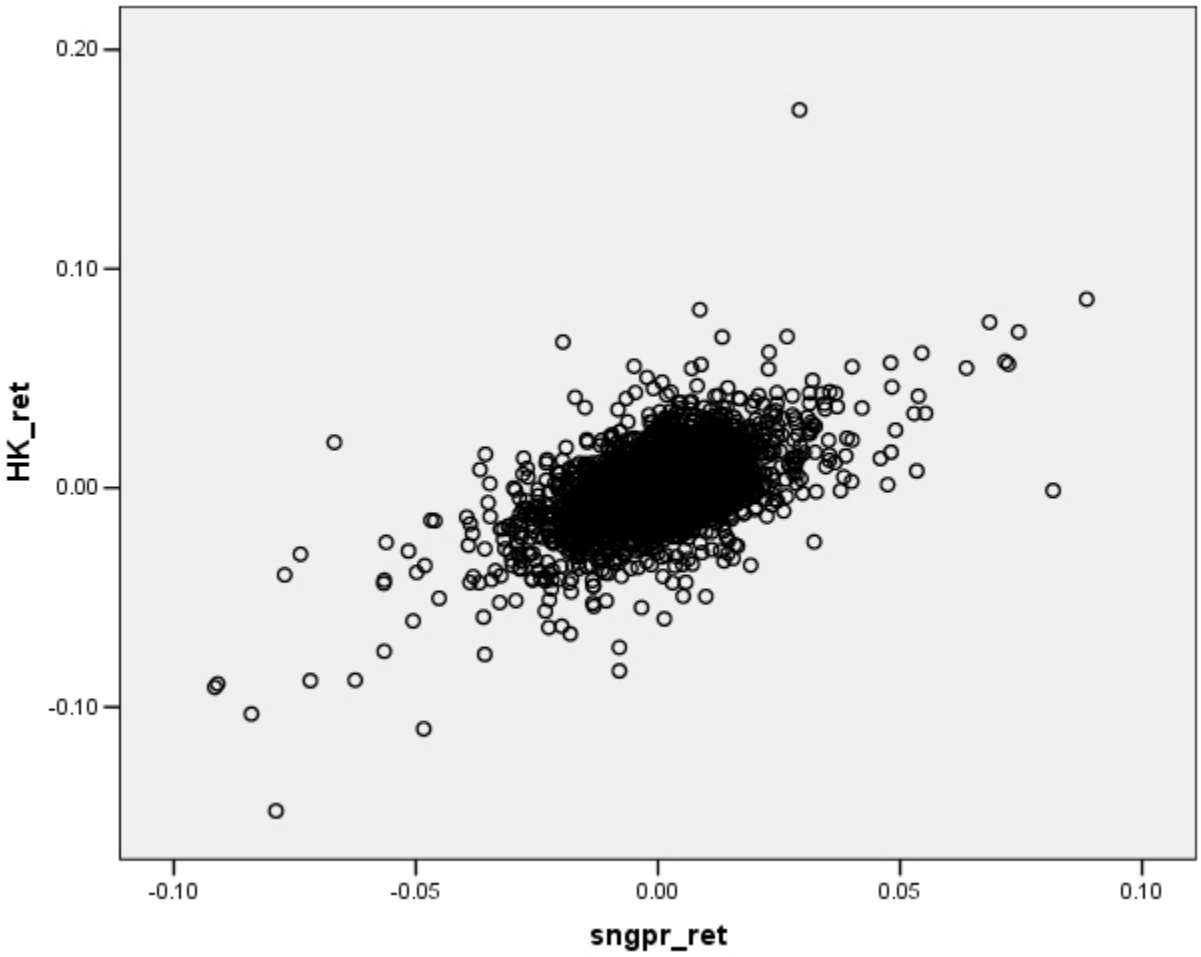
(Let me crush a bit, I learned much of the k-nn stuff from the great book *Doing Data Science* by Cathy O'Neil & Rachel Schutt – get it, read it, love it!)

Jumping into OLS

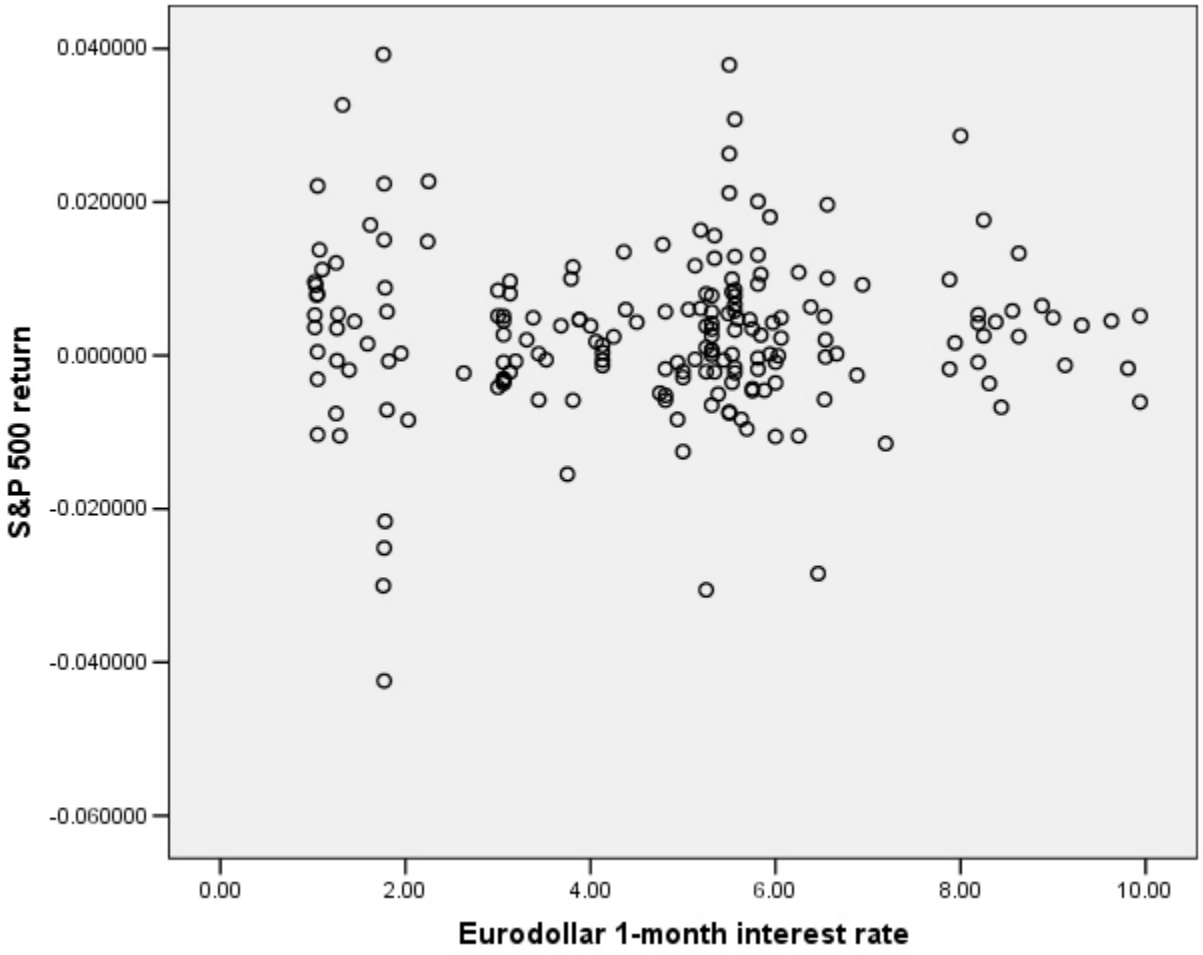
OLS is Ordinary Least Squares, which as the name implies is ordinary, typical, common – something that is widely used in so much analysis.

We are accustomed to looking at graphs that show values of two variables and trying to discern patterns. Consider again these two graphs of financial variables.

This plots the returns of Hong Kong's Hang Seng index against the returns of Singapore's Straits Times index (over the period from Jan 2, 1991 to Jan 31, 2006)



This next graph shows the S&P 500 returns and interest rates (1-month Eurodollar) during 1989-2004.

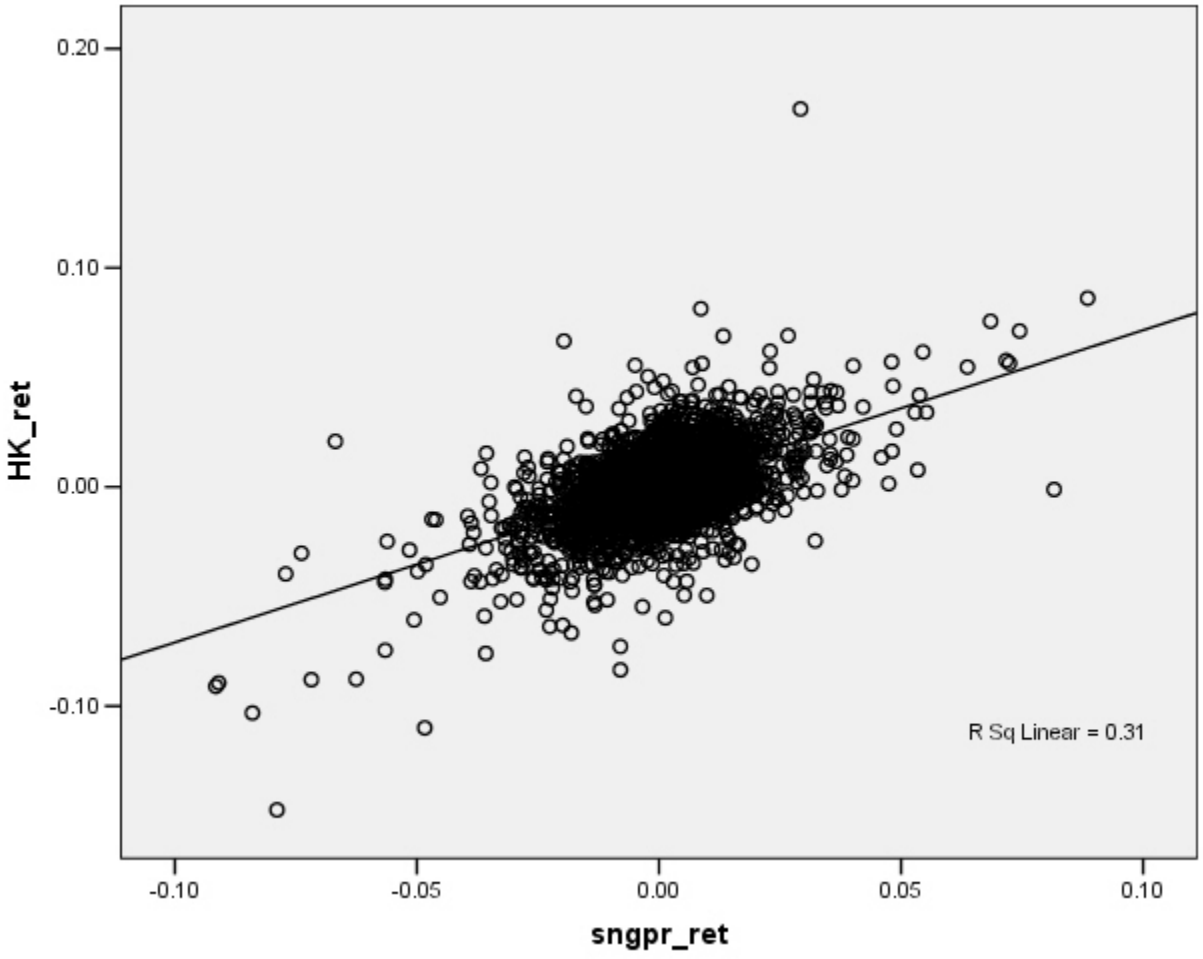


You don't have to be a highly-skilled econometrician to see the difference in the relationships. It would seem reasonable that the Hong Kong and Singapore stock indexes are closely linked while the US stock index is not closely related to interest rates.

So we want to ask, how could we measure these relationships? Since these two graphs are rather extreme cases, how can we distinguish cases in the middle? How can we try to guard against seeing relationships where, in fact, none actually exist? We will consider each of these questions in turn.

How can we measure the relationship?

Facing a graph like the Hong Kong/Singapore stock indexes, we might represent the relationship by drawing a line, something like this:



Now if this line-drawing were done just by hand, just sketching in a line, then different people would sketch different lines, which would be clearly unsatisfactory. What is the process by which we sketch the line?

Typically we want to find a relationship because we want to predict something, to find out that, if I know one variable, then how does this knowledge affect my prediction of some other variable. We call the first variable, the one known at the beginning, X. The variable that we're trying to predict is called Y. So in the example above, the Singapore stock index is X and the Hong Kong index is Y. The line that we would draw in the picture would represent our best guess of what Y would be, given our knowledge about X.

This line is drawn to get the best guess "close to" the actual Y values – where by "close to" we actually minimize the average squared distance. Why square the distance? This is one question which we will return to, again and again; for now the reason is that a squared distance really penalizes the big misses. If I square a small number, I get a bigger number. If I square a big number, I get a HUGE number. (And if I square a number less than one, I get a smaller number.) So minimizing the squared distance will mean that I am willing to make a bunch of small errors in order to reduce a really big error. This is why there is the "LS" in "OLS" -- "Ordinary Least Squares" finds the least squared difference.

A computer can easily calculate a line that minimizes the squared distance between each Y value and the best prediction. There are also formulas for it. (We'll come back to the formulas; put a lightning bolt here to remind us: .)

For a moment consider how powerful this procedure is. A line that represents a relationship between X and Y can be entirely produced by knowing just two numbers: the y-intercept and the slope of the line. In algebra class you probably learned the equation as:

$$Y = mX + b$$

where the slope is m and the y-intercept is b. When $X = 0$ then $Y = b$, which is the value of the line when the line intersects the Y-axis (when X is zero). The y-intercept can be positive or negative or zero. The slope is the value of $\frac{\Delta Y}{\Delta X}$, which tells how much Y changes when X changes by one unit. To find the predicted value of Y at any point we substitute the value of X into the equation.

In econometrics we will typically use a different notation,

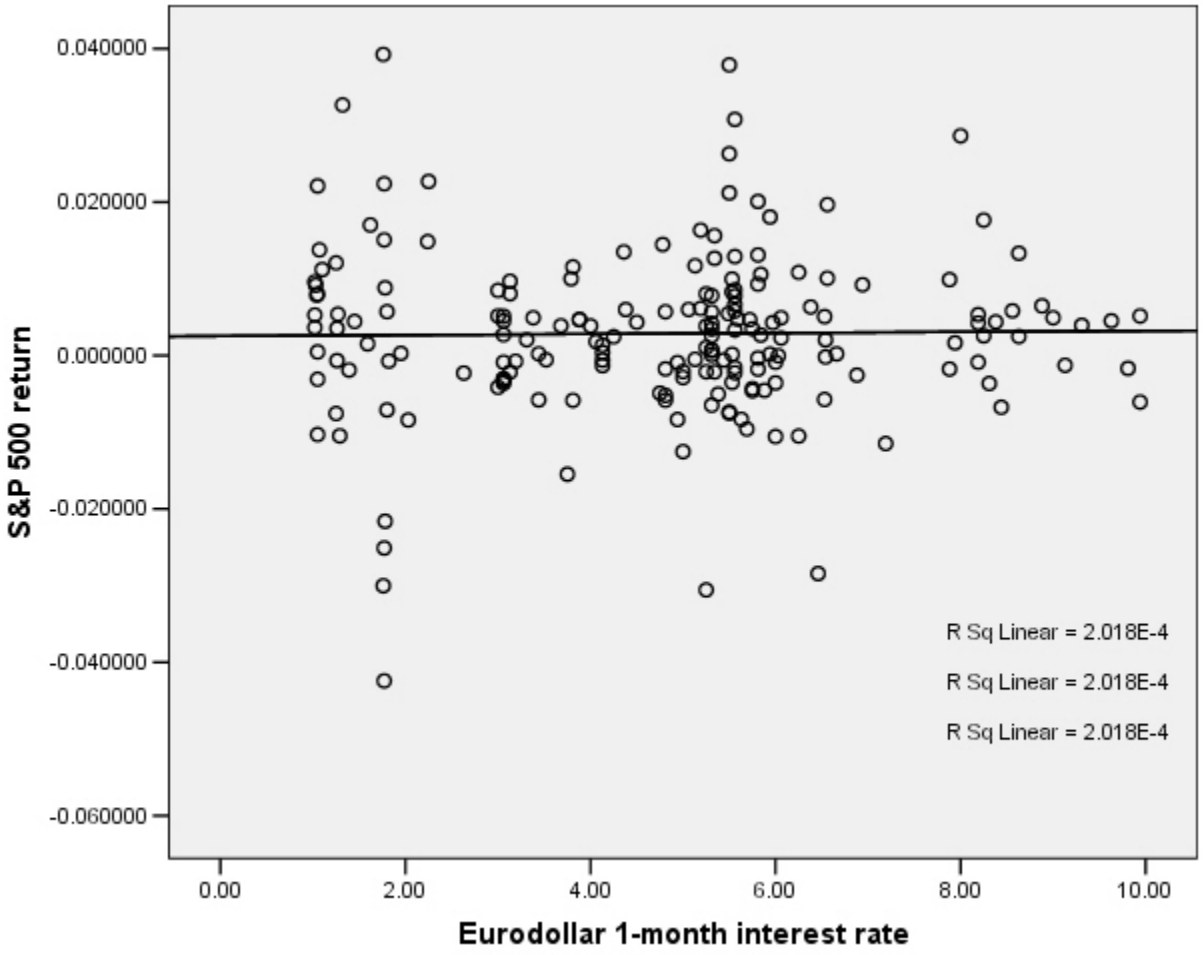
$$Y = \beta_0 + \beta_1 X$$

where now β_0 is the y-intercept and the slope is β_1. (Econometricians looooove Greek letters like beta, get used to it!)

The relationship between X and Y can be positive or negative. Basic economic theory says that we expect that the amount demanded of some item will be a positive function of income and a negative function of price (for a normal good). We can easily have a case where $\beta_1 < 0$.

If X and Y had no systematic relation, then this would imply that $\beta_1 = 0$ (in which case, β_0 is just the mean of Y). In the $\beta_1 = 0$ case, Y takes on higher or lower values independently of what is the level of X.

This is the case for the S&P 500 return and interest rates:



So there does not appear to be any relationship.

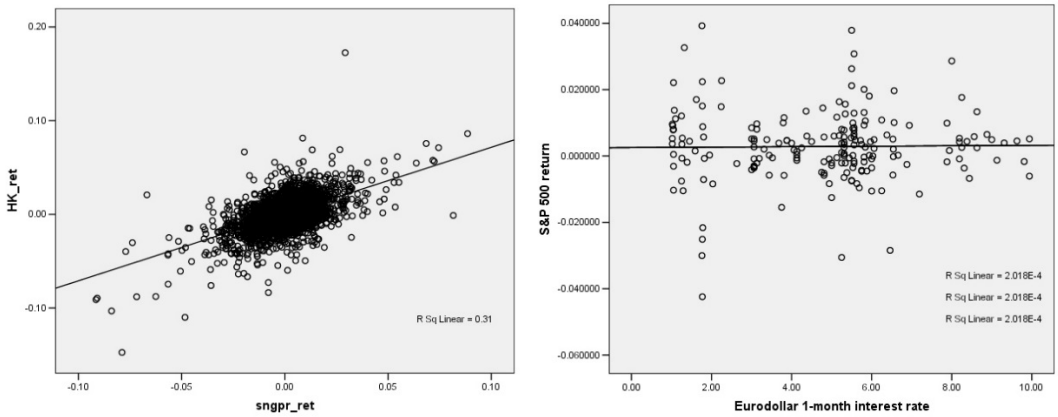
Let's fine up the notation from above a bit more: when we fit a line to the data, we do not always have Y exactly and precisely equal to $\beta_0 + \beta_1 X$. Sometime Y is a bit bigger, sometimes a bit smaller. The difference is an error in the model. So we should actually write $Y = \beta_0 + \beta_1 X + \varepsilon$ where epsilon is the error between the model value of Y and the actual observed value.

Computer programs will easily compute this OLS line; even Excel will do it. When you create an XY (Scatter) chart, then right-click on the data series, "Add Trendline" and choose "Linear" to get the OLS estimates.

Angrist & Pischke distinguish the Conditional Expectation Function as the average value of Y given some X; and OLS is simply the best linear predictor.

How can we distinguish cases in the middle?

Hopefully you've followed along so far, but are currently wondering: How do I tell the difference between the Hong Kong/Singapore case and the S&P500/Interest Rate case? Maybe art historians or literary theorists can put up with having "beauty" as a determinant of excellence, but what is a beautiful line to econometricians?



There are two separate answers here, and it's important that we separate them. Many analyses muddle them up. One answer is simply whether the line tells us useful information. Remember that we are trying to estimate a line in order to persuade (ourselves or someone else) that there is a useful relationship here. And "useful" depends crucially upon the context. Sometimes a variable will have a small but vital relationship; others may have a large but much less useful relation.

This first question, does the line persuade, is always contingent upon the problem at hand; there is no easy answer. You can only learn this by reading other people's analyses and by practicing on your own. It is an art form to be learned, but the second part is science.

The economist Dierdre McCloskey has a simple phrase, "How big is big?" This is influenced by the purpose of the research and the aim of discovering a relation: if we want to control some outcome or want to predict the value of some unknown variable or merely to understand a relationship.

The second question, about the usefulness and persuasiveness of the line, also depends on the relative sizes of the modeled part of Y and the error. Returning to the notation introduced, this means the relative sizes of the predictable part of Y, $\beta_0 + \beta_1 X$, versus the size of ε. As epsilon gets larger relative to the predictable part, the usefulness of the model declines.

The second question, about how to tell how well a line describes data, can be answered directly with statistics, and it can be answered for quite general cases.

How can we try to guard against seeing relationships where, in fact, none actually exist?

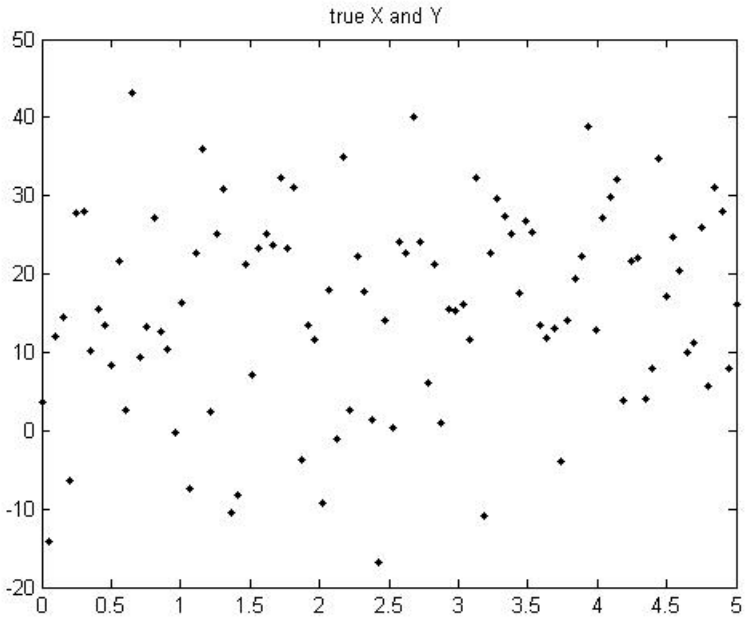
To answer this question we must think like statisticians, do mental handstands, look at the world upside-down.

Remember, the first step in "thinking like a statistician" is to ask, What if there were actually no relationship; zero relationship (so $\beta_1 = 0$)? What would we see?

If there were no relationship then Y would be determined just by random error, unrelated to X. But this does not automatically mean that we would estimate a zero slope for the fitted line. In fact we are highly unlikely to ever estimate a slope of exactly zero. We usually assume that the errors are symmetric, i.e. if the actual value of Y is sometimes above and sometimes below the modeled value, without some oddball skew up or down. So even in a case where there is actually a zero relationship between Y and X, we might see a positive or negative slope.

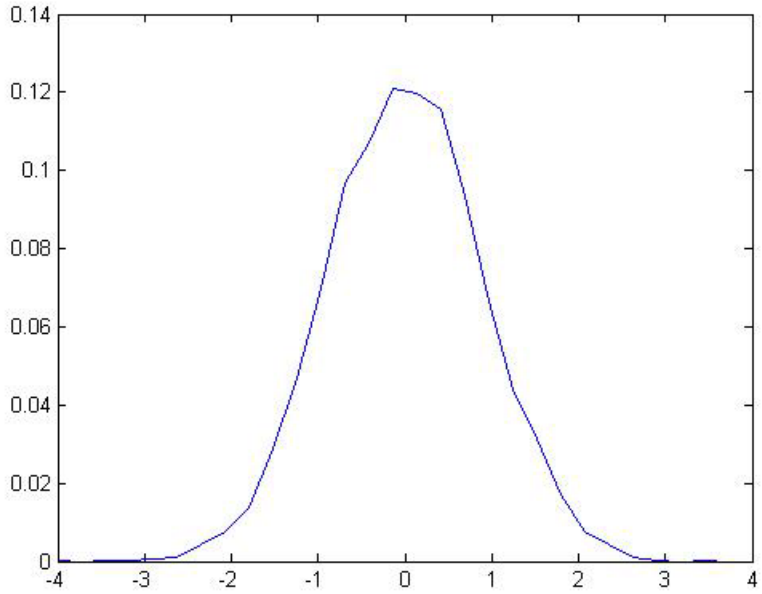
We would hope that these errors in the estimated slope would be small – but, again, "how small is small?"

Let's take another example. Suppose that the true model is Y = 10 + 2X (so $\beta_0 = 10$ and $\beta_1 = 2$). But of course there will be an error; let's consider a case where the error is pretty large. In this case we might see a set of points like this:



When we estimate the slope for those dots, we would find not 2 but, in this case (for this particular set of errors), 1.61813.

Now we consider a rather strange thing: suppose that there were actually zero relationship between X and Y (so that actually $\beta_1 = 0$). Next suppose that, even though there were actually zero relation, we tried to plot a line and so calculated our estimate of β_1. To give an example, we would have the computer calculate some random numbers for X and Y values, then estimate the slope, and we would find 1.45097. Do it again, and we might get 0.36131. Do it 10,000 times (not so crazy, actually – the computer does it in a couple of seconds), and we'd find the following range of values for the estimated slope:



So our estimated slope from the first time, 1.61813, is "pretty far" from zero. How far? The estimated slope is farther than just 659 of those 10,000 tries, which is 6.59%.

So we could say that, if there were actually *no* relationship between X and Y, but we incorrectly estimated a slope, then we'd get something from the range of values shown above. Since we estimated a value of 1.61813, which is farther from zero than just 6.59% if there were actually no relationship, we might say that "there is just a 6.59% chance that X and Y could truly be unrelated but I'd estimate a value of 1.61813."

Now this is a more reasonable measure: "What is the chance that I would see the value, that I've actually got, if there truly were no relationship?" And this percentage chance is relevant and interesting to think about.

This formalization is "hypothesis testing". We have a hypothesis, for example "there is zero relation between X and Y," which we want to test. And we'd like to set down rules for making decisions so that reasonable people can accept a level of evidence as proving that they were wrong. (An example of not accepting evidence: the tobacco companies remain highly skeptical of evidence that there is a relationship between smoking and lung cancer. Despite what most researchers would view as mountains of evidence, the tobacco companies insist that there is some chance that it is all just random. They're right, there is "some chance" – but that chance is, by now, probably something less than 1 in a billion.) Most empirical research uses a value of 5% -- we want to be skeptical enough that there is only a 5% chance that there might really be no relation but we'd see what we saw. So if we went out into the world and did regressions on randomly chosen data, then in 5 out of 100 cases we would think that we had found an actual relation. It's pretty low but we still have to keep in mind that we are fallible, that we will go wrong 5 out of 100 (or 1 in 20) times.

Under some general conditions, the OLS slope coefficient will have a normal distribution -- not a standard normal, though, it doesn't have a mean of zero and a standard deviation of one.

However we can estimate its standard error and then can figure out how likely it is, that the true mean could be zero, but I would still observe that value.

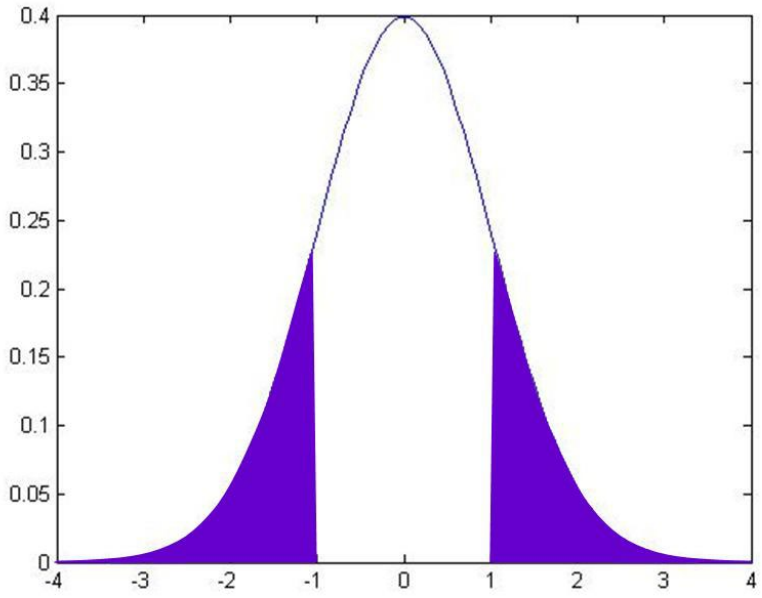
This just takes the observed slope value, call it $\hat{\beta}_1$ (we often put "hats" over the variables to denote that this is the actual observed value), subtract the hypothesized mean of zero, and divide by the standard error:

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)}$$

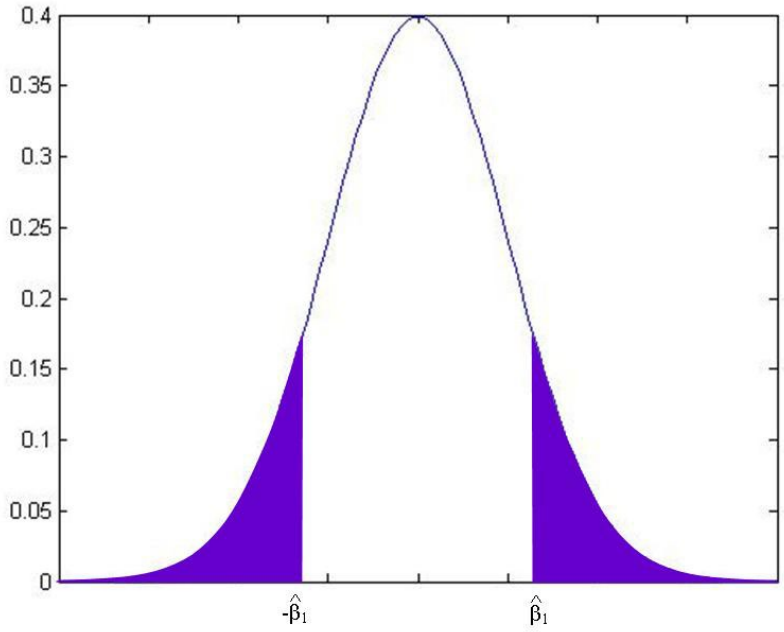
We call this the "t-statistic". When we have a lot of observations, the t-statistic has approximately a standard normal distribution with zero mean and standard deviation of one.

For the careful students, note that the t-statistic actually has a t-distribution, which has a shape that depends on the number of observations used to construct it (the degrees of freedom). When the number of degrees of freedom is more than 30 (which is almost all of the time), the t-distribution is just about the same as a normal distribution. But for smaller values the t-distribution has fatter tails.

The t-statistic allows us to calculate the probability that, if there were actually a zero relationship, I might actually observe a value as extreme as $\hat{\beta}_1$. By convention we look at distance either above or below zero, so we want to know the probability of seeing a value as far from zero as either $\hat{\beta}_1$ or $-\hat{\beta}_1$. If $\hat{\beta}_1$ were equal to 1, then this would be:



while if $\hat{\beta}_1$ were another value, it would be:



From working on the probabilities under the standard normal, you can calculate these areas for any given value of $\hat{\beta}_1$.

In fact, these probabilities are so often needed, that most computer programs calculate them automatically – they're called "p-values". The p-value gives the probability that the true coefficient could be zero but I would still see a number as extreme as the value actually observed. By convention we refer to slopes with a p-value of 0.05 or less (less than 5%) as "statistically significant".

(We can test if coefficients are different from other values than just zero, but for now that is the most common so we focus on it.)

Confidence Intervals for Regression Estimates

There is another way of looking at statistical significance. We just reviewed the procedure of taking the observed value, subtracting off the mean, dividing by the standard error, and then comparing the calculated t-statistic against a standard normal distribution.

But we could do it backwards, too. We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t,$$

and we've just stated that the slope is not statistically significant if:

$|t| < 1.96$.

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se\left(\hat{\beta}_1\right)\right) < \hat{\beta}_1 < 1.96\left(se\left(\hat{\beta}_1\right)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

Details:

- statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is equivalent to the statement that the overall regression explains a statistically significant part of the data variation.
- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart
- There are important assumptions about the regression that must hold, if we are to interpret the estimated coefficients as anything other than within-sample descriptors:
    - X completely specifies the causal factors of Y (nothing omitted)
    - X causes Y in a linear manner
    - errors are normally distributed (for small sample test stats)
    - errors have same variance even at different X (homoskedastic not heteroskedastic)
    - errors are independent of each other
- Because OLS squares the residuals, a few oddball observations can have a large impact on the estimated coefficients, so must explore

Points:

Calculating the OLS Coefficients

The formulas for the OLS coefficients have several different ways of being written. For just one X-variable we can use summation notation (although it's a bit tedious). For more variables the notation gets simpler by using matrix algebra.

The basic problem is to find estimates of β_0 and β_1 to minimize the error in $y_i = \beta_0 + \beta_1 X_i + e_i$.

The OLS coefficients are found from minimizing the sum of squared errors, where each error is defined as $e_i = y_i - \beta_0 - \beta_1 X_i$ so we want to $\min_{\beta_0,\beta_1} \sum_{i=1}^{n} e_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X_i)^2$. If you know basic calculus then you understand that you find the minimum point by taking the derivative with respect to the control variables, so differentiate with respect to β_0 and β_1. After some tedious algebra, find that the minimum value occurs when we use $\hat{\beta}_0$ and $\hat{\beta}_1$, where:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

With some linear algebra, we define the equations as $y = X\beta + e$, where y is a column vector, $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, e is the same, $e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$, X is a matrix with a first column of ones and then columns of each X variable, $X = \begin{bmatrix} 1 & x_1^1 & & x_1^k \\ \vdots & \vdots & \ddots & \\ 1 & x_n^1 & & x_n^k \end{bmatrix}$, where there are k+1 columns, and then $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$. The OLS coefficients are then given as $\hat{\beta} = (X'X)^{-1} X'y$.

But the computer does the calculations so you only need these if you go on to become an econometrician.

To Recap:

- A zero slope for the line is saying that there is no relationship.
- A line has a simple equation, that $Y = \beta_0 + \beta_1 X$
- How can we "best" find a value of β?
- We know that the line will not always fit every point, so we need to be a bit more careful and write that our observed Y values, Y_i ($i=1, \ldots, N$), are related to the X values, X_i, as: $Y_i = \beta_0 + \beta_1 X_i + u_i$. The u_i term is an error – it represents everything that we haven't yet taken into consideration.

- Suppose that we chose values for β_0 and β_1 that minimized the squared values of the errors. This would mean $\min_{\beta_0,\beta_1} \sum_{i=1}^{N} u_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2$. This will generally give us unique values of β (as opposed to the eyeball method, where different people can give different answers).

- The β_0 term is the intercept and the β_1 term is the slope, $\frac{dY}{dX}$.

- These values of β are the Ordinary Least Squares (OLS) estimates. If the Greek letters denote the true (but unknown) parameters that we're trying to estimate, then denote $\hat{\beta}_0$ and $\hat{\beta}_1$ as our estimators that are based on the particular data. We denote $\hat{Y}_i$ as the predicted value of what we would guess Yi would be, given our estimates of β_0 and β_1, so that $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

- There are formulas that help people calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ (rather than just guessing numbers); these are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \text{ and}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \text{ so that } \frac{1}{N}\sum_{i=1}^{N} \hat{Y}_i = \overline{Y} \text{ and } \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$$

Why OLS? It has a variety of desirable properties, if the data being analyzed satisfy some very basic assumptions. Largely because of this (and also because it is quite easy to calculate) it is widely used in many different fields. (The method of least squares was first developed for astronomy.)

- OLS requires some basic assumptions:

  - The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information. But this is why economists look for "natural experiments" where some X is determined by chance outside the ordinary interrelationships.

  - The X and e are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.

  - X_i and u_i have fourth moments. This is technical and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).

- These assumptions are costly; what do they buy us? First, if true then the OLS estimates are distributed normally in large samples. Second, it tells us when to be careful.

- Must distinguish between dependent and independent variables (no simultaneity).

- If these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

- There are formulas that you can use, for calculating the standard errors of the β estimates, however for now there's no need for you to worry about them. The computer will calculate them. (Also note that the textbook uses a more complicated formula than other texts, which covers more general cases. We'll talk about that later.)

Regression in R

To have R do a linear regression, we use the command "lm()" as for example

```
model1 <- lm(Y ~ X1)
summary(model1)
```

This estimates a linear model of $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ and reports estimates of the intercept and slope coefficients.

Regression Details

We'll often form hypotheses about regression coefficients: t-stats, p-values, and confidence intervals – so that's the same basic process as before. Usually two-sided (rarely one-sided).

We will commonly test if the coefficients 'are significant' – i.e. is there evidence in the data that the coefficient is different from zero? This goes back to our original example where we looked at the difference between the Hong Kong/Singapore stock returns and the US stock returns/interest rate. A zero slope is evidence against any relationship – this shows that the best guess of the value of Y does not depend on current information about the level of X. So coefficient estimates that are statistically indistinguishable from zero are not evidence that the particular X variable is useful in prediction.

A hypothesis test of some statistical estimate uses this estimator (call it $\hat{X}$) and the estimator's standard error (denote it as $se_{\hat{X}}$) to test against some null hypothesis value, X_{null}. To make the hypothesis test, form $Z = \dfrac{\hat{X} - X_{null}}{se_{\hat{X}}}$, and – here is the magic! – under certain conditions this Z will have a Standard Normal distribution (or sometimes, if there are few degrees of freedom, a t-distribution; later in more advanced stats courses, some other distribution). The magic happens because if Z has a Standard Normal distribution then this allows me to measure if the estimate of X, $\hat{X}$, is very far away from X_{null}. It's generally tough to specify a common unit that allows me to say sensible things about "how big is big?" without some statistical measure. The p-value of the null hypothesis tells me, "If the null hypothesis were actually true, how likely is it that I would see this $\hat{X}$ value?" A low p-value tells me that it's very unlikely that my hypothesis could be true and yet I'd see the observed values, which is evidence against the null hypothesis.

Often the formula, $Z = \frac{\hat{X} - X_{null}}{se_{\hat{X}}}$, gets simpler when X_{null} is zero, since it is just $Z' = \frac{\hat{X} - 0}{se_{\hat{X}}} = \frac{\hat{X}}{se_{\hat{X}}}$, and this is what R prints out in the regression output labeled as "t".

We know that the standard normal distribution has some important values in it, for example the values that are so extreme, that there is just a 5% chance that we could observe what we saw, yet the true value were actually zero. This 5% critical value is just below 2, at 1.96. So if we find a t-statistic that is bigger than 1.96 (in absolute value) then the slope would be "statistically significant"; if we find a t-statistic that is smaller than 1.96 (in absolute value) then the slope would not be "statistically significant". We can re-write these statements into values of the slope itself instead of the t-statistic.

We know from above that

$$\frac{\hat{\beta}_1 - 0}{se(\beta_1)} = \frac{\hat{\beta}_1}{se(\beta_1)} = t,$$

and we've just stated that the slope is not statistically significant if:

$$|t| < 1.96.$$

This latter statement is equivalent to:

$$-1.96 < t < 1.96$$

Which we can re-write as:

$$-1.96 < \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} < 1.96$$

Which is equivalent to:

$$-1.96\left(se(\hat{\beta}_1)\right) < \hat{\beta}_1 < 1.96\left(se(\hat{\beta}_1)\right)$$

So this gives us a "Confidence Interval" – if we observe a slope within 1.96 standard errors of zero, then the slope is not statistically significant; if we observe a slope farther from zero than 1.96 standard errors, then the slope is statistically significant.

This is called a "95% Confidence Interval" because this shows the range within which the observed values would fall, 95% of the time, if the true value were zero. Different confidence intervals can be calculated with different critical values: a 90% Confidence Interval would need the critical value from the standard normal, so that 90% of the probability is within it (this is 1.64).

OLS is nothing particularly special. The Gauss-Markov Theorem tells us that OLS is **BLUE**: **B**est **L**inear **U**nbiased **E**stimator (and need to assume homoskedasticity). Sounds good, right? Among the linear unbiased estimators, OLS is "best" (defined as minimizing the squared error). But this is like being the best-looking economist – best within a very small and very particular group is not worth much! Nonlinear estimators may be good in various situations, or we might even consider biased estimators.

If X is a binary dummy variable

Sometimes the variable X is a binary variable, a dummy, D_i, equal to either one or zero (for example, female). So the model is $Y_i = \beta_0 + \beta_1 D_i + u_i$ can be expressed as $Y_i = \begin{cases} \beta_0 + \beta_1 + u_i & if\ D_i = 1 \\ \beta_0 + u_i & if\ D_i = 0 \end{cases}$. So this is just saying that Y has mean $\beta_0 + \beta_1$ in some cases and mean β_0 in other cases. So β_1 is interpreted as the difference in mean between the two groups (those with D=1 and those with D=0). Since it is the difference, it doesn't matter which group is specified as 1 and which is 0 – this just allows measurement of the difference between them.

Other 'tricks' of time trends (& functional form)

- If the X-variable is just a linear change [for example, (1,2,3,...25) or (1985, 1986,1987,...2010)] then regressing a Y variable on this is equivalent to taking out a linear trend: the errors are the deviations from this trend. Either the X-variable of (1,2,3,...) or (1985,1986,1987,...) gives the same since the slope coefficient estimates dY/dX and in either case dX=1. There is a difference in the intercept term only.

- If the Y-variable is a log function then the regression is interpreted as explaining percent deviations (since derivative of lnY = dY/Y, the percent change). (So what would a linear trend on a logarithmic form look like?)

- If both Y and X are logs then can interpret the coefficient as the elasticity.

- examine errors to check functional form – e.g. height as a function of age works well for age < 12 but then breaks down

- plots of X vs. both Y and predicted-Y are useful, as are plots of X vs. error.

In addition to the standard errors of the slope and intercept estimators, the regression line itself has a standard error.

A commonly overall assessment of the quality of the regression is the R^2 (displayed by many statistical programs). This is the fraction of the variance in Y that is explained by the model so $0 \leq R^2 \leq 1$. Bigger is usually better, although different models have different expectations (i.e. it's graded on a curve).

Statistical significance for a univariate regression is the same as overall regression significance – if the slope coefficient estimate is statistically significantly different from zero, then this is equivalent to the statement that the overall regression explains a statistically significant part of the data variation.

- Excel calculates OLS both as regression (from Data Analysis TookPak), as just the slope and intercept coefficients (formula values), and from within a chart

Multiple Regression – more than one X variable

Regressing just one variable on another can be helpful and useful (and provides a great graphical intuition) but it doesn't get us very far.

We often know that there are lots of other variables that have influence. How can we keep track of all of these different effects?

Multiple Regression in R

From the standpoint of just using R, there is little difference for the user between a univariate and multivariate linear regression. Again use "`lm()`" but then add a bunch of variables to the model specification, so "`Y ~ X1 + X2 + X3`".

In formulas, model has k explanatory variables for each of $i = (1, 2, \ldots n)$ observations (must have n > k)

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

Each coefficient estimate, notated as $\hat{\beta}_j$, has standardized distribution as t with (n – k) degrees of freedom.

Each coefficient represents the amount by which the y would be expected to change, for a small change in the particular x-variable (i.e. $\beta_j = \frac{\partial y}{\partial x_j}$).

Note that you must be a bit careful specifying the variables. Educational attainment might be coded with a bunch of numbers from 31 to 46 but these numbers have no inherent meaning. If a person graduates high school then their grade coding changes from 38 to 39 but this must be coded with a dummy variable. If a person moves from New York to North Dakota then this increases their state code from 36 to 38; this is not the same change as would occur for someone moving from North Dakota to Oklahoma (40) nor is it half of the change as would occur for someone moving from New York to North Carolina (37). Each state needs a dummy variable. These X-variables are not continuous.

A multivariate regression can control for all of the different changes to focus on each item individually.

Example

- BMI as function of Age, Sleep, Education
- `model_OLS <- lm(X_BMI5 ~ SLEPTIM1 + EDUCA + X_AGEG5YR, data = brfss22)`
- `summary(model_OLS)`

<more details in lecture video>

Take the output a piece at a time. First it confirms what model you had called (useful when you go back later, after you've run lots of regressions). Next it gives a summary of the residuals,

$$\varepsilon_i = y_i - \hat{y} = y_i - \left(\widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i}\right)$$

These can be called at any point with "*`residuals(model3)`*" so the output is simply from "`summary(residuals(model3))`". The mean is not reported here since the model constrains the mean

of the residuals to zero. The fitted values, $\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_2} x_{2,i} + \cdots + \widehat{\beta_k} x_{k,i}$, can be called as `fitted.values(model3)`. You can plot these.

Then R reports the coefficients, standard errors, t-statistics, and p-values for each term in the model. The coefficients and standard errors are calculated by the estimation routine. The t-statistic is the ratio of the coefficient estimate divided by the standard error, $t = \dfrac{\hat{\beta}}{se\left(\hat{\beta}\right)}$. The p-value is the area in the tails of a t-distribution (with degrees of freedom as shown on bottom line) beyond the t-statistic. The command, "`coefficients(model3)`", accesses the coefficient values.

At the bottom of the R summary it shows the R-squared, the standard error of the residual (which is basically the same as `sd(residuals(model3))`), and the F-statistic, which is another measure of how well the model fits.

You should be able to calculate the t-statistic and p-value from the coefficient estimates and standard errors by yourself (the next homework will give you some chances to practice that).

You should also be able to calculate confidence intervals, although R can do that for you as well, with for example, `confint(model3,level = 0.95)`.

R will also produce lots of plots, simply with `plot(model3)`, which gives lots of plots in sequence – you can pick off particular ones with `plot(model3, which = 3)` that will give the 3rd plot. (The plots indicate that this might not be a great model.)

You can get an Analysis of Variance (ANOVA) with `anova(model3)`. For now don't worry about the details of the output except to the final row of figures, labeled "`Residuals`". This gives one of the most important bits of information about the model: how big are the residuals? Remember that's the whole point of the OLS estimator – it minimizes the (squared) residuals. So this gives you the value of the sum of squared residuals.

There is a final detail, that we use `interval = "confidence"` if the x-values to be predicted are inside the values estimated, and `interval = "prediction"` if the x-values are outside.

Statistical Significance

Statistical significance of coefficient estimates is the same when we look at individual coefficients but more complicated for multiple coefficients: we can ask whether a group of variables are jointly significant, which takes a more complicated test. We can even ask if all of the slope coefficients together are statistically significant.

For a univariate regression, if the single slope coefficient is statistically significant then the overall regression is as well (the F statistic is the square of the t-stat in that case).

The difference between the overall regression fit and the significance of any particular estimate is that a hypothesis test of one particular coefficient tests if that parameter is zero; is $\beta_i = 0$? This uses the t-statistic $t = \frac{\hat{\beta}}{se\left(\hat{\beta}\right)}$ and compares it to a t distribution. The test of the regression significance tests if ALL of the slope coefficients are simultaneously zero; if $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_K = 0$. The latter is much more restrictive. (See Chapter 7 of Stock & Watson.)

It is often sensible to make joint tests of regression coefficients, for example with a group of dummy variables. If we have a set of dummies for education levels, it is strange to think of omitting just one or two; it is more reasonable to ask whether education measures (overall) are statistically significant. We might also want to know if individual coefficients are equal to each other (e.g. to ask if going to college, without getting any degree, is really different from the estimate for just a high school diploma.

To do this in R, there is a package, `linearHypothesis` (part of the package, `car`, Companion to Applied Regression, which is auto-loaded by `AER` package). But the commands shouldn't obscure the simple basic point: we evaluate variables based on how well they fit in the model.

To consider the question of whether a set of variables is statistically significant, we basically are just looking at how big is the error (the Sum of Squared Errors) with and without those variables. In general adding more variables to the model can never make the errors bigger (can never increase the Sum of Squared Errors) – basically this is a statement that the Marginal Benefit of more variables can never be negative. But profit maximization requires that we balance Marginal Benefit against Marginal Cost – what is the marginal cost of adding more variables? Statistical significance is one measure of profitability in this sense.

If adding new predictors makes the error "a lot" smaller, then those predictors are jointly statistically significant. The essence of statistical testing is just finding a good metric for "a lot".

Note that we can only properly make comparisons within models – it doesn't make much sense to look across models. If I have a model of the fraction of income spent on food, and another model of the level of income, it is difficulty to sensibly pose a question like, "in which model is education more important?" It would be like asking who scored more points per game, Shaq or Jeter? – you can ask the question but it's difficult to interpret in a sensible way.

But within a model we can make comparisons and many of them come down to asking, how much smaller are the errors? (Did the Sum of Squared Errors fall by a lot?) Sometimes it is easiest to just estimate the model twice, with or without the variables of interest, and look at how much the Sum of Squared Errors (from ANOVA in R) fell. But once you get some experience, you'll appreciate `linearHypothesis`.

Finally note that "statistically significant" is different from "important". Suppose you have some Y-values ranging from 100 – 1000, but you notice that a particular X value is associated with the first decimal

value. When X has one value, the first decimal is .2; when X has another value the first decimal is 0.7. There are a lot of reasons that could be the case. This could be an interesting pattern and this could tell us subtle things about the world. But a 0.5 difference, among values ranging over 3 digits, is really tiny! A hypothesis of statistical significance could duly tell you that the X-value is significant (it is a good indicator of whether the outcome is yyy.2 or yyy.7). But depending on the question you're asking, that could be unimportant.

Why do we always leave out a dummy variable? Multicollinearity.

- OLS basic assumptions:
  - The conditional distribution of u_i given X_i has a mean of zero. This is a complicated way of saying something very basic: I have no additional information outside of the model, which would allow me to make better guesses. It can also be expressed as implying a zero correlation between X_i and u_i. We will work up to other methods that incorporate additional information.
  - The X and errors are i.i.d. This is often not precisely true; on the other hand it might be roughly right, and it gives us a place to start.
  - X and errors don't have values that are "too extreme." This is technical (about existence of fourth moments) and broadly true, whenever the X and Y data have a limit on the amount of variation, although there might be particular circumstances where it is questionable (sometimes in finance).
- So if these are true then the OLS are unbiased and consistent. So $E\left[\hat{\beta}_0\right] = \beta_0$ and $E\left[\hat{\beta}_1\right] = \beta_1$. The normal distribution, as the sample gets large, allows us to make hypothesis tests about the values of the betas. In particular, if you look back to the "eyeball" data at the beginning, you will recall that a zero value for the slope, β_1, is important. It implies no relationship between the variables. So we will commonly test the estimated values of β against a null hypothesis that they are zero.

Factors in R

R has a shortcut for lots of dummy variables – some variables are labeled as factors. Try them in your regressions.

But remember the math behind. A factor coding education might have levels for if the person has a highschool diploma, if they have some college, if they have a college degree. Those are actually a bundle of yes/no questions, coded in usual Boolean manner that 1 is yes and zero is no.

	Is highest level of education...		
Education factor	Highschool	Some college	College degree
Highschool	1	0	0
Some college	0	1	0
College degree	0	0	1

Testing if All the New Variable Coefficients are Zero

You're wondering how to tell if all of these new variables are worthwhile. Simple: Hypothesis Testing! There are various formulas, some more complicated, but for the case of homoskedasticity the formula is relatively simple.

Why any formula at all – why not look at the t-tests individually? Because the individual t-tests are asking if each individual coefficient is zero, not if it is zero and others as well are also zero. That would be a stronger test.

To assess any model, we look at how well in predicts and what it misses. To measure how much a group of variables contributes to the regression, we look at the residual values – how much is still unexplained, after the various models? And since this is OL**S**, we look at the **squared** residuals. R outputs the Sum of Squares for the Residuals in the ANOVA. We compare the sum of squares from the two models and see how much it has gone down with the extra variables. A big decrease indicates that the new variables are doing good work. And how do we know, how big is "big"? Compare it to some given distribution, in this case the F distribution. Basically we look at the percent change in the sum of squares, so something like:

$$F \approx \frac{SSR_0 - SSR_1}{SSR_1}$$

with the wavy equals sign to show that we're not quite done. Note that model 0 is the original model and model 1 is the model with the additional regressors, which will have a smaller residual (so this F can never be negative).

To get from approximately equal to an equals sign, we need to make it a bit like an elasticity – what is the percent change in the number of variables in the model? Suppose that we have N observations and that the original model has K variables, to which we're considering adding Q more observations. Then the original model has (N – K – 1) degrees of freedom [that "1" is for the constant term] while the new model has (N – K – Q – 1) degrees of freedom, so the difference is Q. So the percent change in degrees of freedom is $\frac{Q}{N-K-Q-1}$. Then the full formula for the F test is

$$F = \left(\frac{SSR_0 - SSR_1}{SSR_1}\right) \Big/ \left(\frac{Q}{N-K-Q-1}\right).$$

Which is, admittedly, fugly. But we know its distribution, it's F with (Q, N-K-Q-1) degrees of freedom – the F-distribution has 2 sets of degrees of freedom. Calculate that F, then use R to find pf(F, df1 =Q, df2 = (N-K-Q-1)) (or Excel to calculate FDIST(F,Q,N-K-Q-1)), to find a p-value for the test. If the p-value is less than 5%, reject the null hypothesis.

Usually you will have the computer spit out the results for you. In R, `anova(model1, model2)` or else `linearHypothesis()` as we did before.

Factors Interacting

aka moderators, intersectionality, etc...

Nonlinear Regression

(more properly, **How to Jam Nonlinearities into a Linear Regression**)

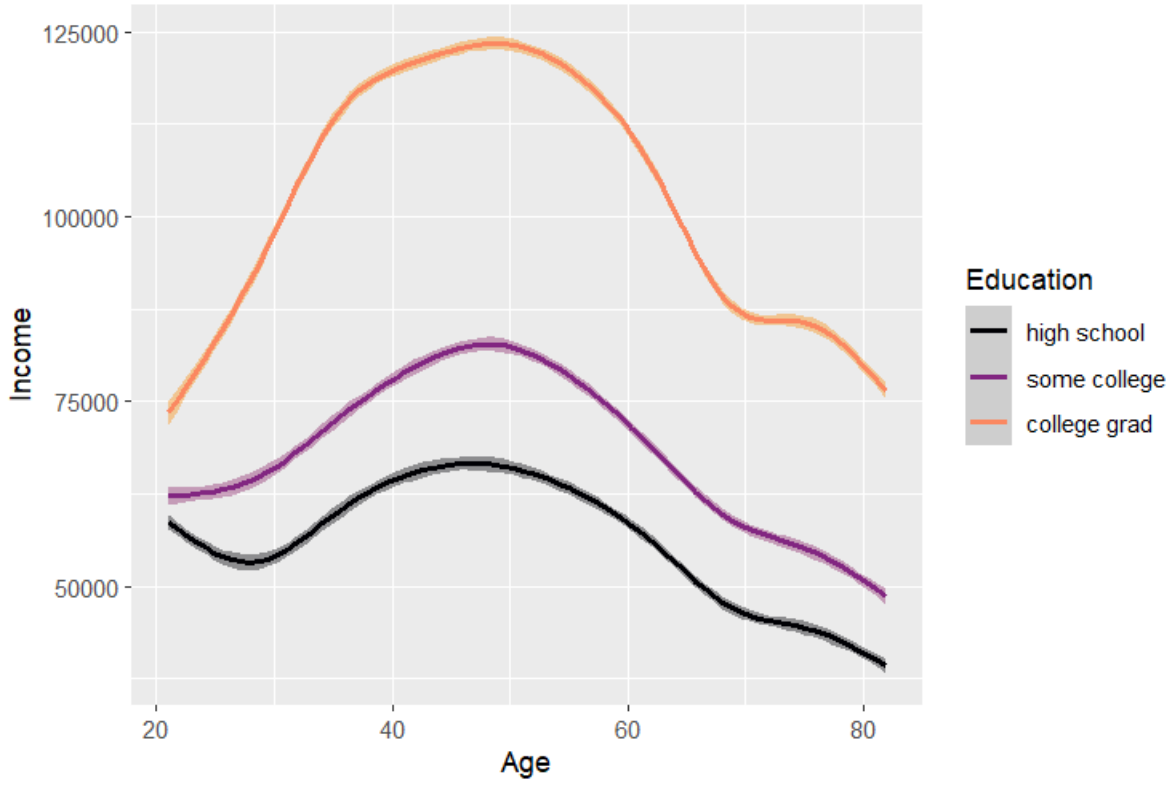
- $X, X^2, X^3, \ldots X^r$
- ln(X), ln(Y), both ln(Y) & ln(X)
- interactions of dummy/continuous
- interactions of continuous variables

There are many examples of, and reasons for, nonlinearity. In fact we can think that the most general case is nonlinearity and a linear functional form is just a convenient simplification which is sometimes useful. But sometimes the simplification has a high price. For example, my kids believed that age and height are closely related – which is true for their sample (i.e. mostly kids of a young age, for whom there is a tight relationship, plus 2 parents who are aged and tall). If my sample were all children then that might be a decent simplification; if my sample were adults then that's lousy.

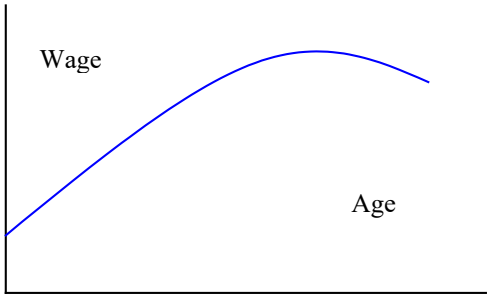
The usual justification for a linear regression is that, for any differentiable function, the Taylor Theorem delivers a linear function as being a close approximation – but this is only within a neighborhood. We need to work to get a good approximation.

Nonlinear terms

Why is our regression linear? This is mostly convenience, and we can easily add non-linear terms such as Age^2, if we think that the typical age/wage profile is not linear. For example, ggplot showed this relationship between age and income for different educational levels:



A first approximation might be to estimate those as being like a part of a parabola,



So the regression would be:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \ldots + \varepsilon_i$$

(where the term "..." indicates "other stuff" that should be in the regression).

As we remember from calculus,

$$\frac{dWage}{dAge} = \beta_1 + \beta_2 \cdot 2 \cdot Age$$

so that the extra "boost" in wage from another birthday might fall as the person gets older, and even turn negative if the estimate of $\beta_2 < 0$ (a bit of algebra can solve for the top of the hill by finding the Age that sets $\frac{dWage}{dAge} = 0$).

We can add higher-order effects as well, maybe Age^3 and Age^4 terms, which can trace out some complicated wage/age profiles. However we need to be careful of "overfitting" – adding more explanatory variables will never lower the R^2.

Logarithms

Similarly can specify X or Y as ln(X) and/or ln(Y).

(You also need to figure out how to work with observations where Y=0 since ln(0) doesn't give good results. Dropping those observations might be OK or might not, it depends.)

If Y is in logs and D is a dummy variable, then the coefficient on the dummy variable is just the percent change when D switches from zero to one.

So the choice of whether to specify Y as levels or logs is equivalent to asking whether dummy variables are better specified as having a constant level effect (i.e. women make $10,000 less than men) or having a percent change effect (women make 25% less than men). As usual there may be no general answer that one or the other is always right.

Dummies

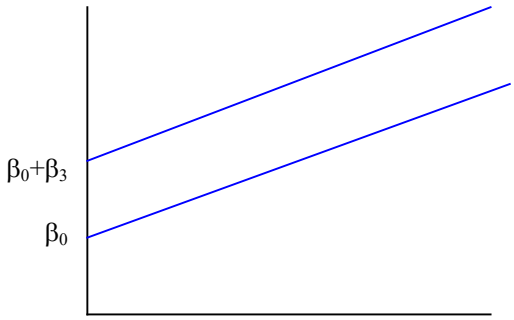
Recall our discussion of dummy variables, that take values of just 0 or 1, which we'll represent as D_i. Since, unlike the continuous variable Age, D takes just two values, it represents a shift of the constant term. So the regression,

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + u_i$$

The equation could be also written as

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ \beta_0 + \beta_3 + \beta_1 Age_i + u_i & for\ D = 1 \end{cases}.$$

These show that people with D=0 have intercept of just β_0, while those with D=1 have intercept equal to $\beta_0 + \beta_3$. Graphically, this is:



We need not assume that the β_3 term is positive – if it were negative, it would just shift the line downward. We *do* however assume that the *rate* at which age increases wages is the same for both genders – the lines are parallel.

Dummy Variables Interacting with Other Explanatory Variables

The assumption about parallel lines with the same slopes can be modified by adding interaction terms: define a variable as the product of the dummy times age, so the regression is

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_3 D_i + \beta_4 D_i Age_i + u_i$$

or

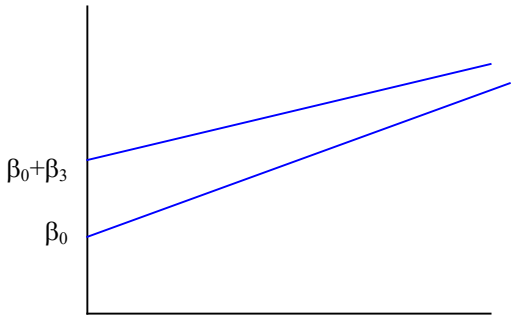
$$Wage_i = (\beta_0 + \beta_3 D_i) + (\beta_1 + \beta_4 D_i) Age_i + u_i$$

or

$$Wage_i = \begin{cases} \beta_0 + \beta_1 Age_i + u_i & for\ D = 0 \\ (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age_i + u_i & for\ D = 1 \end{cases}$$

so that, for those with D=0, as before $\frac{\Delta Wage}{\Delta Age}$=$\beta_1$ but for those with D=1, $\frac{\Delta Wage}{\Delta Age} = \beta_1 + \beta_4$.

Graphically,

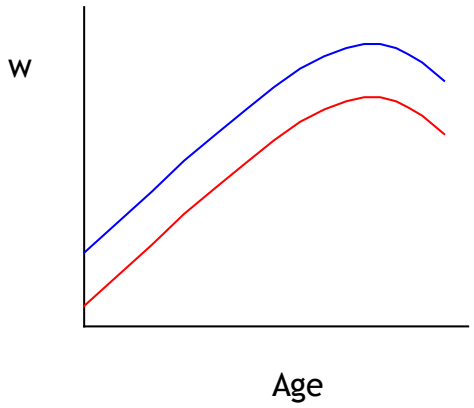


so now the intercepts and slopes are different.

So we might wonder if men and women have a similar wage-age profile. We could fit a number of possible specifications that are variations of our basic model that wage depends on age and age-squared. The first possible variation is simply that:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 D_i + u_i ,$$

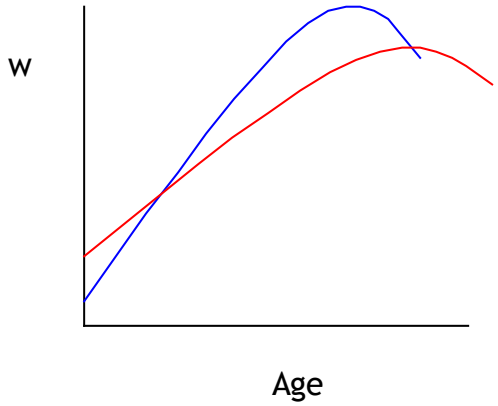
which allows the wage profile lines to have different intercept-values but otherwise to be parallel (the same hump point where wages have their maximum value), as shown by this graph:



The next variation would be to allow the lines to have different slopes as well as different intercepts:

$$Wage_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2$$
$$+\beta_3 D_i + \beta_4 D_i Age_i + \beta_5 D_i Age_i^2 + u_i$$

which allows the two groups to have different-shaped wage-age profiles, as in this graph:



(The wage-age profiles might intersect or they might not – it depends on the sample data.)

We can look at this alternately, that for those with D=0,

$$wage = \beta_0 + \beta_1 Age + \beta_2 Age^2$$
$$\frac{dWage}{dAge} = \beta_1 + 2\beta_2 Age$$

so the extreme value of Age (where $\frac{dWage}{dAge} = 0$) is $\frac{-\beta_1}{2\beta_2}$.

While for those with D=1,

$$Wage = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \beta_3 + \beta_4 Age + \beta_5 Age^2$$
$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) Age + (\beta_2 + \beta_5) Age^2$$
$$\frac{dWage}{dAge} = (\beta_1 + \beta_4) + 2(\beta_2 + \beta_5) Age$$

so the extreme value of Age (where $\frac{dWage}{dAge} = 0$) is $\frac{-(\beta_1 + \beta_3)}{2(\beta_2 + \beta_4)}$. Or write the general value, for both cases, as $\frac{-(\beta_1 + \beta_3 D)}{2(\beta_2 + \beta_4 D)}$ where D is 0 or 1.

This specification, with a dummy variable multiplying each term: the constant and all the explanatory variables, is equivalent to running two separate regressions: one for men and one for women:

$$D = 0$$
$$Wage_i = \beta_0^{male} + \beta_1^{male} Age_i + \beta_2^{male} Age_i^2 + u_i$$
$$D = 1$$
$$Wage_i = \beta_0^{female} + \beta_1^{female} Age_i + \beta_2^{female} Age_i^2 + e_i$$

Where the new coefficients are related to the old by the identities: $\beta_0^{female} = \beta_0 + \beta_3$, $\beta_1^{female} = \beta_1 + \beta_4$, and $\beta_2^{female} = \beta_2 + \beta_5$. Sometimes breaking up the regressions is easier, if there are large datasets and many interactions.

Note that it would be very weird (and difficult to justify) to have an interaction of the dummy with the Age term but not with Age-squared or vice versa. Why would we want to assume that, say, men and women have different linear effects but the same squared effect?

Interactions with R

It is very easy to do interactions with R, maybe too easy so that you can forget what it all means.

It can be difficult to unpack the meaning all of the interaction terms. The regression creates dummy variables for educational classifications, showing that people with progressively higher educational qualifications get more money. But women get less at each rung: the coefficients on female interacted with education are negative. So for instance a male with an associate's degree is predicted to make about $20,700 more than a male without even a high school diploma, but a woman with an associate's degree gets $8400 less than the man – so her net premium for the associate's degree is (20,700 – 8400) = 12,300.

We can create a table showing the net values, like this (also setting Age = 30),

	Intercept+(Age=30)	HS	Some Coll	AS	Bach	Adv Deg
male	24494	10570	20178	20737	44536	79607
female	-6494	-6501	-9045	-8391	-15904	-30213
difference						
net	18001	4068	11133	12347	28632	49394

So in equations this says that

$$wage = \beta_0 + \beta_1 Age + \beta_2 Female + \beta_3 EducHS + \beta_4 EducSomeC + \beta_5 EducAS + \beta_6 EducBach + \beta_7 EducAdv + \cdots \{other\ stuff\} + \ldots + \gamma_1 Female * EducHS + \gamma_2 Female * EducSomeC + \gamma_3 Female * EducAS + \gamma_4 Female * EducBach + \gamma_5 Female * EducAdv + \varepsilon$$

Then the predicted values are, say for a 30-year-old female with an associate's degree,

$$\widehat{wage} = \beta_0 + \beta_1(Age = 30) + \beta_2(Female = 1) + \beta_3(EducHS = 0) + \beta_4(EducSomeC = 0) + \beta_5(EducAS = 1) + \beta_6(EducBach = 0) + \beta_7(EducAdv = 0) + \cdots \{other\ stuff\} + \ldots + \gamma_1(Female = 1) * (EducHS = 0) + \gamma_2(Female = 1) * (EducSomeC = 0) + \gamma_3(Female = 1) * (EducAS = 1) + \gamma_4(Female = 1) * (EducBach = 0) + \gamma_5(Female = 1) * (EducAdv = 0)$$

Which looks ferociously complicated but multiplying by zero drops many of the terms

$$\widehat{wage} = \beta_0 + \beta_1(Age = 30) + \beta_2(Female = 1) + \beta_3(EducHS = 0) + \beta_4(EducSomeC = 0) + \beta_5(EducAS = 1) + \beta_6(EducBach = 0) + \beta_7(EducAdv = 0) + \cdots \{other\ stuff\} + \ldots + \gamma_1(Female = 1) * (EducHS = 0) + \gamma_2(Female = 1) * (EducSomeC = 0) + \gamma_3(Female = 1) * (EducAS = 1) + \gamma_4(Female = 1) * (EducBach = 0) + \gamma_5(Female = 1) * (EducAdv = 0)$$

From staring at the wage penalties, you might also conclude that it looks somewhat multiplicative, that the wage penalty for females is around 35%-40% for all of the terms involving college. This might motivate a log specification (which is usually preferred in the literature, I'm just passing over it here in order not to overwhelm with ornamentation).

You might next look at gender/marital status interactions, or education/race/ethnicity interactions – there is no reason you can't do interactions upon interactions. They get complicated but just write out the various interactions in long equation format to help remember what is what. Just don't be a monkey about interpreting and understanding all of the interactions. The limit on how many interactions comes since as you take finer and finer cuts, you're essentially looking at group means where the numbers in each group get smaller and smaller. So you can do state-level factors interacted with gender and education, and probably get a decent estimate of how the wages of women-with-associates-in-NY compares with wages of women-with-associates-in-Cali, but worse estimate of women-with-associates-in-Wyoming or some other empty state where nobody lives. Multi-level models (later) try to deal with this problem.

Binary Dependent Variable Models

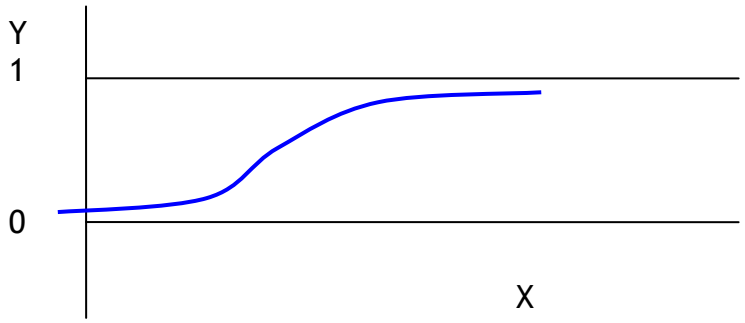
- Sometimes our dependent variable is continuous, like a measurement of a person's age; sometimes it is just a "yes" or "no" answer to a simple question. A "Yes/No" answer can be coded as just a 1 (for Yes) or a 0 (a zero for "no"). These zero/one variables are called dummy variables or **binary** variables. Sometimes the dependent variable can have a range of discrete values ("How many children do you have?" "Which train do you take to work?") – in this case we have a discrete variable. The binary and continuous variables can be seen as opposite ends of a spectrum.

- We want to explore models where our dependent variable takes on discrete values; we'll start with just binary variables. For example, we might want to ask what factors influence a person to go to college, to have health insurance, or to look for a job; to have a credit card or get a mortgage; what factors influence a firm to go bankrupt; etc.

- Linear Models such as OLS have some problems. These imply predicted values of Y that are greater than one or less than zero. They also have advantages! You should be able to do both http://marcfbellemare.com/wordpress/8951

- Interpret our prediction of Y as being the probability that the Y variable will take a value of one. (Note: remember which value codes to one and which to zero – there is no necessary reason, for example, for us to code Y=1 if a person has health insurance; we could just as easily define Y=1 if a person is uninsured. The mathematics doesn't change but the interpretation does!)

- want to somehow "bend" the predicted Y-value so that the prediction of Y never goes above 1 or below zero, something like:



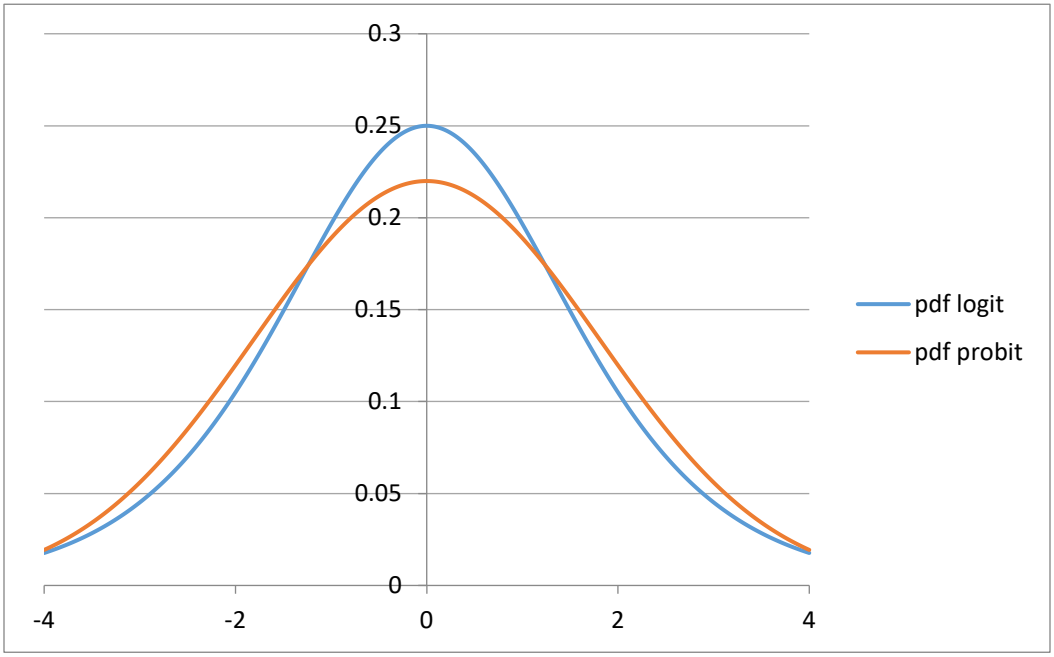
- Probit Model

- $\Pr(Y=1|X)=\Phi(\beta_0+\beta_1X_1+\beta_2X_2)$ where $\Phi(\square)$ is the cdf of the standard normal
- $\frac{\Delta \Pr}{\Delta X}$ is not constant

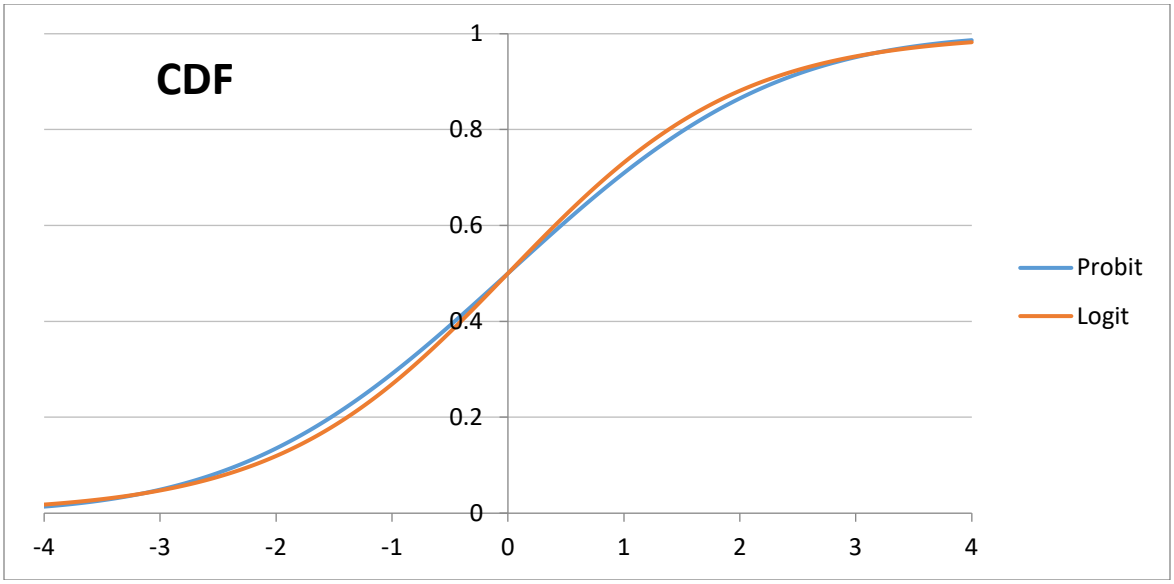
- Logit Model
  - $\Pr(Y=1|X)=F(\beta_0+\beta_1X_1+\beta_2X_2)$, where $F(z)=\frac{1}{1+e^{-z}}$
  - $\frac{\Delta \Pr}{\Delta X}$ is not constant

- differences (Excel sheet:compare_probit_logit.xls)

Clearly the differences are rather small; it is rare that we might have a serious theoretical justification for one specification rather than the other.



(Note that the logit function given above has standard error of $\frac{\pi}{\sqrt{3}}$ so in the plots I scaled the probit by this factor).



- Measures of Fit
    - no single measure is adequate; many have been proposed
    - What probability should be used as "hit"? If the model says there is a 90% chance of Y=1, and it truly is equal to one, then that is reasonable to count as a correct prediction. But many measures use 50% as the cutoff. Tradeoff of false positives versus false negatives – loss function might well be asymmetric.

	actually = 1	*actually = 0*
Predicted = 1	*Hooray!*	*sad*
Predicted = 0	*sad (maybe sadder?)*	*Hooray!*

Probit/Logit in R

For a logit estimation, just

```
regn_logit1 <- glm(Y ~ X1 + X2, family = binomial, data = data1)
```

for a probit estimation

```
regn_probit1 <- glm(Y ~ X1 + X2, family = binomial (link = 'probit'), data = data1)
```

Then the estimation results from "`summary()`" should be familiar. The interpretation is also essentially unchanged: look at the individual t-statistics (formed by dividing coefficient estimates from standard errors) then get a p-value from that.

In addition to looking at effects of particular X-variables, we are interested in looking at predictive accuracy – but note that this is likely to vary depending on your project so the results I'm going to show here are particular to this analysis. You would have to carefully take a look at your own model predictions. Also would want to check different sub-groups – is predictive accuracy substantially better or worse for particular groups? That might be a signal that the simple dummies are not adequately capturing the variation.

Also note that the code as given treats either miss (whether actually true and predict false, or actually false and predict true) as equally bad. In many applications this is not the case! Depending on the purpose of the model, false negatives and false positives could have different costs.

- Details of estimation

- recall that OLS just gives a convenient formula for finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize the sum $\sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}\right)\right)^2$. If we didn't know the formulas we could just have a computer pick values until it found the ones that made that squared term the smallest.

- similarly a probit or logit coefficient estimates are finding the values of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ that minimize $\sum_{i=1}^{n}\left(Y_i - f\left(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}\right)\right)^2$, whether the $f(\square)$ function is a normal c.d.f. or a logit c.d.f.

- Maximum Likelihood (ML) is a more sophisticated way to find these coefficient estimates – better than just guessing randomly.

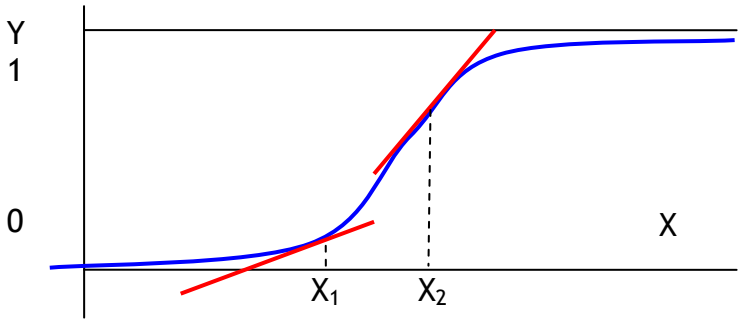
- For example the likelihood of any particular value from a normal distribution is the p.d.f., $\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$. If we have 2 independent observations, X_1, X_2 from a distribution that is known to be normally distributed with variance of 1 (to keep the math easy) then the joint likelihood is $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_2-\mu)^2}$. We want to find a value of μ that maximizes that function. This is an ugly function but we could note that any value of μ that maximizes the natural log of that function will also maximize the function itself (since $\ln(\square)$ is monotonic) so we take logs to get $\ln\left(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(X_1-\mu)^2 - \frac{1}{2}(X_2-\mu)^2$. Take the derivative with respect to μ and set it equal to zero to get $(X_1-\mu)+(X_2-\mu)=0$ so that $\mu = \frac{(X_1+X_2)}{2}$. You should be able to see that starting with n observations would get us $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ so the average is also the maximum-likelihood estimator. A maximum-likelihood estimator could be similarly derived in cases where we don't know the variance (interestingly, that ML estimator of the standard error divides by *n* not *(n – 1)* so it is biased but consistent).

- Maximizing the likelihood of the probit model is one or two steps more complicated but not different conceptually. Having a likelihood function with a first and second derivative makes finding a maximum much easier than the random hunt.

Properly Interpreting Coefficient Estimates:

Since the slope, $\frac{\Delta Y}{\Delta X} = \frac{\Delta \Pr}{\Delta X}$, the change in probability per change in X-variable, is always changing, the simple coefficients of the linear model cannot be interpreted as the slope, as we did in the OLS model. (Just like when we added a squared term, the interpretation of the slope got more complicated.)

Return to the picture to make this clearer:



The slope at X_1 is rather low; the slope at X_2 is much steeper.

The effect of the coefficients now interacts with all of the other variables in the model: The biggest difference is toward the middle.

Multi-Level Modeling

After Fixed Effects, we can generalize to Multi-Level Modeling (much of my explanation is based on the excellent book, *Data Analysis Using Regression and Multilevel/Hierarchical Models*, by Andrew Gelman & Jennifer Hill). From the wage regressions based on CPS data that we were using, we can consider adding information about the person's occupation (the data gives a rough grouping of people into about 20 occupations). You've probably done a version of this regression in your head, if you've ever read someone's job title and tried to figure out how much she makes.

There are a few ways to use the occupation data. One way is to ignore it, to not use it – which is what we were doing when we left it out of the regression. Everyone started from the same value. Gelman & Hill call this the "pooling" estimator since it pools everyone together. Another way would be to put in fixed effects for each occupation, letting each vary as needed – every occupation has a different intercept term, starting from a different value. This is "no-pooling." This puts no constraints at all on what the intercepts might be – some high, some low, some way far afield. A multilevel model imposes a model on how those intercepts vary: usually that they have a normal distribution with a central mean and variance. The math to define the estimator gets a bit more complicated, but we let the computer worry about that. But it's basically a weighted average of the "pooled" and "no-pooled" estimates, where the number of people reporting being in that particular group give the weights. So groups with a lot of members get nearly that "no-pooled" estimate, while a group with few members would be estimated to be like the larger group.

So in this example, the pooling case has wages of person i in industry j explained as $w_{i,j} = \alpha + \beta X_{i,j} + e_{i,j}$ (where the X includes all the rest of the variables, lumped together). The no-pooling case has $w_{i,j} = \alpha_j + \beta X_{i,j} + e_{i,j}$ so the intercept varies by industry, j. The multilevel case has $w_{i,j} = \alpha_0 + \alpha_{[j]} + \beta X_{i,j} + e_{i,j}$ but $\alpha_{[j]} \sim N(0, \sigma_\alpha)$.

With just a single level (like Occupation) this doesn't seem like a big thing, but if we want to define a lot of levels (Occupation, Industry, State or even City) then this gets more important. Instead of estimating a separate parameter for each level, we can estimate just overall parameters – and levels with only a small number of observations will be partially pooled.

Once we decide we want to do such a thing, the remaining question is, "how?" With R it's easy, just `lmer()` instead of `lm()`. [y ~ (1 + z | group) + x + z + x:z]

```
modelmm1 <- lmer(WSAL_VAL ~ as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm1)

modelmm2 <- lmer(WSAL_VAL ~ Age + female + AfAm + Asian + Amindian + race_oth
               + Hispanic + educ_hs + educ_smcoll + educ_as + educ_bach + educ_adv
               + married + divwidsep + union_m + veteran + immigrant + immig2gen
              + as.factor(A_DTOCC) + (1 | as.factor(A_DTOCC)), dat_use)
summary(modelmm2)
```

In these cases we can compute the Intra-Class Correlation (ICC) which is the ratio of the variance in the groups (σ_α) to the total variance, so $\frac{\sigma_\alpha}{\sigma_\alpha+\sigma_\epsilon}$. Kind of like R^2, this goes from zero to one and is graded on a curve. It tells how important the within-group variation is, relative to the total variation.

Of course the next step would be to expand these coefficient estimates to be for slope as well as intercept – something like $w_{i,j} = \alpha_0 + \alpha_{[j]} + \left(\beta_0 + \beta_{[j]}\right)X_{i,j} + e_{i,j}$. Multilevel modeling is a growing trend within econometrics.