Glossary of Statistical Terms

Confidence interval:

A pollster says something like: The candidate has a 20-point lead with margin of error equal to +/- 2 points.
In statistics, it is common to form a confidence interval by multiplying standard error [See STANDARD ERROR] by 2 and then adding/subtracting it from the estimated statistic. So in this example 16 to 24 points represents a 95% confidence interval: we can say with 95% confidence that the true lead is between 16 and 24 points..

Note: You do not have to create a **95%** confidence interval even though it is customary to do so.

One standard error: You could throw caution to the wind and add and subtract the standard error (in this case +/-2) **without multiplying** it. Then, you could say with 68% confidence that the true lead was between 18 and 20 points.

Three standard errors: You could also be super conservative and create a 99.7% confidence interval. A 99.7% confidence interval is created by multiplying the standard error by **three** and then adding and subtracting it from the result. In this case, you could say with 99.7% confidence that the true lead is between 14 and 26.

Why 68% or 95% or 99.7%? See NORMAL CURVE.

Correlation:

Let's say you line up all the kids in a class according to height. You then line them up again according to weight. The correlation indicates the extent to which students will occupy the same spot in the height line that they occupied in the weight line. Correlations run from +1 to -1. A correlation of +1 means that the two lines are exactly the same: The tallest kids are also the heaviest. He is at the front of both lines. A correlation of -1 indicates that the lines are the same except they run in reverse order from one another: The tallest kids is the lightest and the shortest kid is the heaviest. The tallest kid is at the back of the weight line but he is at the front of the height line.

What do people mean when they say correlation does not necessarily equal causation?

Now let's say we first line up kids in the class by the amount of time they spend texting. Then, we line them up by their reading test scores. There is a positive correlation, i.e. kids standing near the front of the texting line are also near the front of the math score line.

So can we conclude that texting causes higher math scores? Should the teacher start assigning students to text each other as math homework?

Not necessarily. We need more information.

It could easily be the case that the students who text the most have the fanciest cell phones, and these students also come from wealthier families. Students from wealthier families also tend to score higher on reading tests. In this example family wealth would confound the correlation between texting and reading scores. Texting may not be the cause of those higher scores even though the two are correlated.

Mean, median, mode:

Mean: The classic, traditional average. As in: I wrote three stories on Monday, five stories on Tuesday, one story on Wednesday, one story Thursday and two stories Friday. So my average daily story count is 3+5+1+1+2=10. Ten divided by 5 (the number of days in my work week) equals two. In other words, two is the average number of stories I wrote each day I worked.

Median: I line up my daily story count in size order. 1,1,2,3,5. The number in the very center of my line is the median. (I like to remember that the highway has a median and it is in the very center of the road.) So the median is 2. What if I had worked a cops shift on Sunday and had written six stories? How would I determine the median when I have two numbers in the center, 2 and 3?
Here are my story counts lined up in a row:
1,1,2,3, 5,6.
All I do is average together the two central numbers. 2+3=5. 5 divided by 2 (because I am averaging 2 numbers) is 2.5.

Why not just use the mean? My publisher has established a daily byline quota but I don't really feel very motivated this week. So I sit back and relax and don't write any stories until Friday, when I suddenly crank out five crappy stories.
Which is a better indicator of my overall work effort this week? The mean of 1? (0+0+0+0+5 divided by 5= 1.) Or the median of 0? (0,0,0,0,5)?

More generally, medians are useful any time you have some very large or very small measurements that are going to provide a skewed picture of what's really going on. A good example is income. The Warren Buffetts of this world skew mean income calculations, so we often use the median.

Mode: Mode is the poor, neglected third wheel of the trio, the one you learned in school and probably immediately forgot. Mode means the most common number in a series. (I remember it because a mode is a trend and when something is trendy, you see a lot of it.) So, the return to the example above, if my daily story count is 0,0,0,0, 5, the mode is 0 because 0 is the most common number.

When would I ever use the mode? For some data indicators, neither the mean nor the median is really meaningful. For instance, let's say I have 10 boys in a class and 15 girls. What would it really mean for me to calculate the "average" gender or the median gender? However, the mode (15) is meaningful because it tells me I have more girls than boys.

Measurement error:

Virtually every measurement is imperfect. If the measurement were taken again, it would differ, even if only slightly. Even if we agree that a test is written to measure exactly what we want to know about a student, it will measure this imperfectly. The "measurement error" can be thought of as the difference between the results of the measurement (e.g. the test score) and the true value of interest (e.g., a student's knowledge of 7th grade math).

With tests, it is possible to compute a "standard error of measurement" that can be associated with each student's score. For instance, let's say you score a 500 on a 1,000-point test. If the standard error of measurement for that score is 50 points, that means that if you took the test again, you would likely get a score ranging from 450 to 550. Usually, standard errors of measurement are bigger for very low scores and smaller for very high scores.

Why should I care about measurement error?

Consider a test that divides people into categories. (Proficient/not proficient, pass/no pass). What if a student gets a 500? Let's say the standard error of measurement for that score is 50, meaning the student would probably score between 400 and 600 if he took the test again. Specifically, there's a 95% chance that the 'true' performance is between 400 and 600. What if the cut off for graduating from high school or passing on to the next grade level is a 501? What should be done?

Now consider the cases of very high or very low scores: Here's a real example from the 2011 third-grade Colorado Student Assessment Program reading exam. Scale scores on the exam ranged from 150 points to 795 points. The standard error for the exam as a whole was 26 points.

But the standard error of measurement for a student who got the top score of 795 was 56 points. The standard error of measurement for the student who got the bottom score of 150 was 233 points.

How should these scores be interpreted?

But why would different test scores have different measurement errors?

Individual questions on a test can provide more information about students who ultimately end up with certain test scores. For example, a test question may do a really good job at distinguishing between a student who will ultimately get the top score on a test versus a student who will ultimately get a slightly lower score on the test as a whole. The top scoring student will probably get the answer correct. The slightly lower-scoring student will ultimately get the answer wrong.

But this question does not provide much information about the student who gets a very low score versus the student gets the lowest score. That's because both of these students will probably get the answer wrong.

Usually on a test, you will have more questions that differentiate among people in the middle ranges because most test takers will fall into those ranges. Also, with tests that use cut scores, it is important to have lots of questions that differentiate among people on the cusp of the cut points for each of the categories (e.g Are you proficient of partially proficient? Did you pass or fail the high school exit exam?) Such tests often contain fewer questions that differentiate among people who end up with scores that are far from the cut points. So the test provides less overall information about these students.

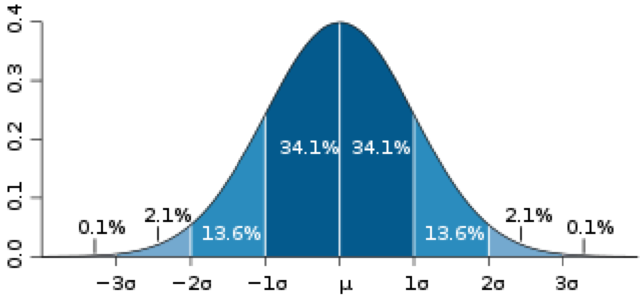
When we have lots of information about people who end up with a test score similar to yours, then the standard error for your test score is low. When we have very little information distinguishing you from your neighbor who ends up with a slightly higher or lower score, then the standard error for your test score is high.

Normal Curve:

It turns out that lots of data are distributed in a pattern called a “normal” or “bell curve.” (E.g., as with height or SAT scores.) In a normal curve, most of the data points are close to the average but a few are way above average or way below average.

Sometimes people make the mistake of assuming that we arranged students’ test scores or other data into a normal curve. We did not. It just happens that there are a lot of situations in which a lot of people are average but some are way above or way below the average.

Statisticians like to use normally-distributed data because a lot of their tools assume that the data set has this distribution. When data are not normally distributed, these tools may sometimes spit out misleading results.



The graph above shows normally distributed data. Normal curves have some very nice properties.

When data are normally distributed, about **68%** of the data (e.g. the number of people who earned each test score) is one standard deviation higher or lower than the average.

About **95%** of data are within two standard deviations of the average.

About **99.7%** of data are within three standard deviations of the average.

Normal curves are kind of like Barbie: These measurements [68%, 95%, 99.7%] are always the same, regardless of whether you are looking at IQ or height or widgets or weevils.

Why care about the normal curve?

Normal curves can help identify significant or unusual results. They can be used to assign students’ test scores to percentiles that compare the scores to others in the distribution. They can even help identify cheaters, as is evident from the excerpt from USA Today:

USA TODAY used a methodology widely recognized by mathematicians, psychometricians and testing companies. It compared year-to-year changes in test scores and singled out grades within schools for which gains were 3 standard deviations or more from the average statewide *gain on that test. In layman's language, that means the students in that grade showed greater improvement than 99.9% of their classmates statewide.*

P values and Statistical Significance:

Interview with a p value.

Q: Who are you?
A: I'm glad you asked because a lot of people don't really get me. My official bio says that I am "the possibility of getting a result at least this extreme IF the null hypothesis is correct." A lot of people find that kind of intimidating. Who even knows what the null hypothesis is? I like to think of myself as proof by contradiction. I tell people, I am the probability that you got your result even though no real effect exists. So, most people want me to be really, really small.

Q: Tell me about a typical day in your life.
A: Well, the other day a researcher I know was testing a vaccine against procrastination. She randomly divided her subjects into two pools: The treatment group got the vaccine. The control group got salt water. She got really good results: Kids in the treatment group procrastinated a lot less than kids in the control group. My value, the p value, was .04.

Q: Your value, the p value, was .04. What does that mean?
A: Good question. That means that if the vaccine was truly no more effective than giving salt water to both groups, and if she did the study 100 times, she would probably get the same fantastic results that she found only four times. There's only a 4% chance that the great results she got this one time were by chance. It's pretty unlikely the salt water would give her that result. So she's allowed to advertise her results as statistically significant.

Q: Why .04 ? What if you had been .06?
A: By convention, we say that when I am .05 or less, the result is "statistically significant." When I go on a diet and get my weight down to .01 or less, I am "highly statistically significant." I know it's random, but those cutoffs are just what everyone has agreed upon and so that's what researchers use. When you see researchers trumpeting results that are "significant" because I am less than .10, then you know they're reaching.

Q: Can you give us some tips on dealing with you and your kind?
A: I am really, really sensitive to sample size. So sometimes, because of the way I am calculated, when the data set is really big, I get tired -- and small differences end up highly statistically significant. Sorry about that.

Q: What about when the data set is really small?
A: Oh. Yeah. As I said, very sensitive to sample size--it makes me break out in hives. So with a small data set, I might tell you that your results are not statistically significant but only because the sample size was not large enough to detect an effect.

Q: But I thought you were above all that kind of stuff! If researchers can't depend on you, how are they supposed to know whether their results are meaningful?!
A: Look, I'm not some magic wand that you people can just wave over their analyses. I think maybe I've been over-hyped. All I ever promised is to tell researchers the probability that

they got that result, but there is nothing going on with their data. Read my contract! I'm not anybody's data maid. A researcher has to look at her own results and all the context surrounding them to decide whether her differences are meaningful in the real world. That's called substantive significance or practical significance, and it's not my department.

Randomized Controlled Trials (a.k.a. "RCT" or "Experiments"):

In a randomized controlled trial, subjects are randomly assigned to a treatment group and a control group. The treatment group gets the intervention (a medicine, a reading program). The control group does not. If possible, the control group should get a "placebo" that resembles the treatment but is neutral. That way, you can tell whether people are responding to the treatment versus the idea that they are getting a treatment. For instance, if I were to give the treatment group a new medicine, I might give the control a sugar pill. This placebo would help me distinguish between reactions that were caused by the medicine versus reactions that were caused by people thinking they got the medicine. As you can imagine, in the social sciences, it can be difficult to come up with an appropriate placebo.

When researchers call something an "experiment," they are referring to a randomized controlled trial. They tend to get annoyed if you call any other research design an "experiment." For instance, researchers would get irritated if you wrote this sentence: "The school district is trying an experiment in which every single student gets an Ipad." That's not an experiment. There is no randomly-assigned control group of kids who *didn't* get an Ipad.

A "double-blind experiment" is a type of randomized controlled trial in which neither the subjects nor the evaluators know who is in the treatment group and who is in the control group. Again, this can be difficult in the social sciences. If your daughter suddenly has 10 people in her class while your son has 40, it's hard not to notice.

You might have heard a randomized controlled trial called the "gold standard." That's because the researcher has randomly assigned subjects to treatment or to control groups. This makes it very likely that differences between the groups are due to either chance (which can be assessed using statistics) or to the treatment. When an experimental treatment "works", you can say that it "caused" an effect. By contrast, when the researcher has not assigned subjects to treatment and control groups, she generally will only go so far as to say that an effect is "associated" with a certain cause.

"Cause" is a loaded term for researchers. You will see a lot of linguistic gymnastics around this hesitancy to use the word "cause" in any context other than a randomized controlled experiment. Researchers who did not use an experimental design may get mad if you write that they found that something "caused" something else--e.g. that Ipad use caused an increase in math scores in that earlier example where everyone in the district got an Ipad.

If the randomized controlled trial is so fantastic, then why don't we use it every time in educational research? Here are some reasons:

* *Practical considerations:* It can be difficult to get people to volunteer themselves for an experiment, much less volunteer their children, especially if the treatment is particularly desirable or particularly risky. It can also be difficult to convince school officials to permit a researcher to, for instance, randomly assign kids to a class or teachers to a reading program. Once the experiment is under way, participants may try to switch from treatment to control or vice versa. (E.g. "My kid was assigned to the large class last year that was the control group but I'm going to sneak him into the small class/treatment group next year.") Or they might drop out entirely if they decide the experiment is harmful or worthless. When classes within the

same school are randomized into treatment and control groups, effects may bleed over from the treatment group to the control group because kids and teachers in one group interact with kids and teachers in the other. It can take a lot of time, money, staff and attention to initiate and maintain this design in the real world.

**Ethical concerns:* Some of the effects we care most deeply about in the social sciences and even in medicine are not conditions that would be ethical to randomly assign to certain groups. We cannot randomly assign kids to poor and rich parents. Would it be ethical to randomly assign some teenagers to drink a liter of vodka a day so we could determine how chronic alcohol use affected their bodies and their minds? For this very reason, some of the most well-accepted scientific conclusions have been proven without randomized controlled trials. For example, the conclusion that smoking causes lung cancer.

**Type of Research Question:* Randomized controlled experiments are fantastic for answering the question: Does A cause B? However, there are many other questions of interest that do not necessarily require an experiment or are not suited to an experiment. For instance, what are the characteristics of this population of learners? Why or how does this school reform model lead to these results?

* *Generalization:* If randomized controlled experiments do not suffer problems such as attrition, they can have very strong “internal validity.” But they often have weak “external validity.” These terms are defined below. But imagine an oversubscribed charter school that holds a lottery. The losers are studied as a control group . (winners = treatment group). The study is carried out, and the findings show a benefit. We can say that winning the lottery “caused” the benefit (strong internal validity). But we can’t say much about the charter on the other side of the city, which is a different model and which isn’t oversubscribed (weak external validity).

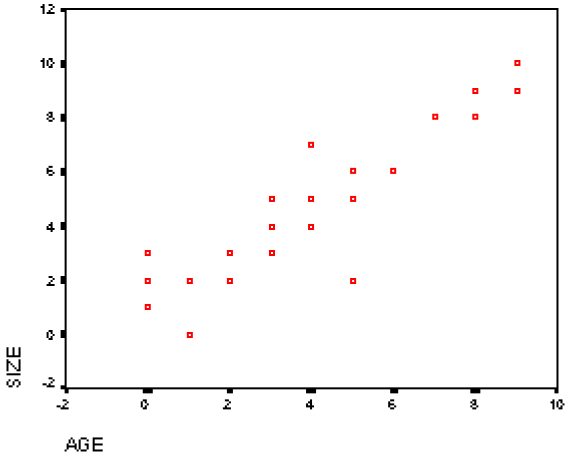
Regression Analysis:

In regression analysis, you are trying to predict (at least one) variable using at least one other variable.
There is a big secret about regression analysis that I will explain in a few paragraphs. But first:

The variable that you are trying to predict has several a.k.a.’s, including “outcome variable,” and “dependent variable.”

The variables you are using for the prediction also go by several names, including “independent variable,” “control variables” and “explanatory variables.”

Let’s say you work at the Humane Society. Someone has brought in a bunch of young shih tzus...again. You want to give the people who adopt this latest group of puppies an idea of how much bigger they can expect the dogs to get in the next few months. Since the second littler of dogs came from the same parents as the first litter, you believe that the growth trajectories will be fairly similar. So, based upon your last load of shih tzus, you create this scatter plot of age (in months) versus weight (in pounds).

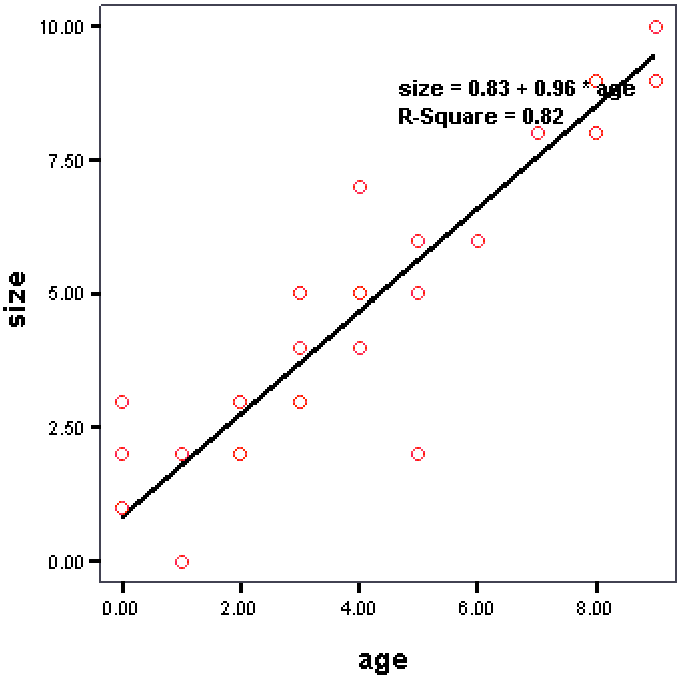


Each red dot= 1 puppy.

On the horizontal or X axis, is the age in months.

On the vertical or Y axis is the size in pounds.

A regression line is simply the line that will best connect the dots. By “best” I mean it is drawn with maximal closeness to each dot. The graph below shows the dogs, their ages in months, their weights in pounds and that regression line.



Here’s the big secret: The regression line is not some big, complex mystery that only a few people learn about in advanced graduate courses. Yes, regression analysis can get complicated. But the regression line is determined by the same old formula you and millions of other kids were taught in your high school or middle school math class. Remember this?

y=mx+ b

Y is the thing you are predicting. In this case, it is weight.

I like to think of **B** as a sort of base level/ base camp/beginning point. It is what Y equals when X is 0.

Here, it is a puppy's age at birth, i.e. when he is 0 months old. At 0 months, the puppy weighs less than 1 pound--or .83 of a pound.

m tells us how many units that **x** increases when **y** increases by one unit.
Or, in English: When a puppy gets 1 month older, how much weight does the regression line predict she will gain?
The answer is that she will gain .96 of a pound or nearly one pound per month.

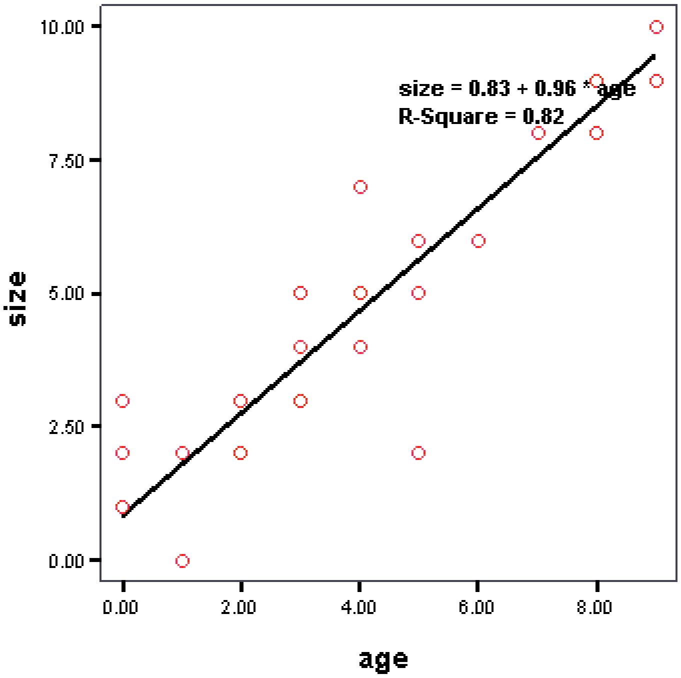
Here's how I would translate this entire equation into English:

The regression line predicts that a puppy will weigh .83 of a pound at birth and will gain nearly 1 additional pound (.93) each month.

These numbers that define the regression line (.83 and .96) are called parameter estimates.

Here's another cool thing about a regression line: It can suggest which individuals (dogs, schools, etc) are performing above or below what we predicted.

Let's look again at the graph of the puppies, their ages and their weights. (I have re-pasted the same graph below) What is going on with that dog who is five months old and still weighs 2 and a half pounds? It looks like he might be the runt of his litter. Or maybe he just needs more food. And what about that puppy who weighed more than 2 and a half pounds at birth. Is she maybe a Great Dane mix?



The distance from the regression line to the individual point of data (e.g.. the individual dog) is called the **residual**. A dog below the regression line is light for his age. A dog above the regression line is heavy for his age. The farther you are from the regression line, the worse of a job the regression line is doing at predicting your individual case.

Value-added scores and residuals

And now, if you understand that, you know the basics of how value-added scores are computed: A value-added score is a residual. A positive value-added score means you are doing better than predicted--i.e. you are above the regression line. A negative value-added score means you are doing worse than predicted--i.e. you are below the regression line. Yes, it's a little more complex than that. Teachers' value-added scores are generally calculated using a particular form of regression analysis called hierarchical linear modeling. This accounts for the fact that we are interested in the collective performance of the students assigned to each teacher. But the basic concept is the same. You are still looking at the distance from the regression line, i.e the residual.

R-squared and variability

In the above example of the dogs, we are doing a pretty good job of predicting their weights. Most of the dogs are close to the regression line. The average residual is not that big. For the most part, we can say that variability in age explains most of the variability in weight: If you're trying to estimate the weight of one of these shih tzu puppies, you can feel comfortable saying that it would help to know the dog's age.

There is actually a measurement that tells us how much of the variability we have explained with our regression line. For instance, how important is age versus amount of food eaten or ..some other factor. This measurement is called an R-squared. In this case, the R-squared is .82. This

means that our explanatory variable (age) has explained 82% of the variability of weight.
If age had explained all of the variability in the dogs' weights, the R-squared would have been 1. All the dots representing the dogs would have been perfectly lined up on the regression line.
If age had explained none of the variability in weight, the dots representing the dogs would have been nowhere near the regression line. R-squared would have been 0.

What is a "good" R squared?

As with many questions in statistics, the answer is, it depends. Let's say your study "explains" 50 percent of the variability in test scores -- e.g., R-squared = .50. But past studies have explained 75% (R squared=.75). You might ask yourself, what am I missing? On the other hand, let's say we know that 80% in the variability of a test score is "explained" by something that is impractical or impossible to change -- e.g. gender. If your reading intervention explained 10%, then it might be worthwhile, even revolutionary.

Why care about R-squared?

Some of you have probably heard researchers say something like: "60% in the variability of test scores can be explained by out-of-school factors." Or: "When it comes to student achievement, teachers are the most important in-school factor. " What the researchers are referring to is the R-squared. What they are saying is that...if you have a regression line that is only predicted using in-school factors (.e.g school funding, teacher quality), you will have a lot of kids who are really far from that regression line because there are a lot of out-of-school factors you are not accounting for (e.g, parental income level, parental education level). Your regression line might not be in the right place. Here's the rub: even if your regression line is in the wrong place because you are failing to account for all sorts of important variables (or because you don't have information about these variables), you can still generate parameter estimates -- i.e. you can still get this type of result:

For each percentage point that the college attendance rate increases, the mean ACT score of the school increases by a corresponding .15 of a point.

Sounds pretty official, right? But how much of the variability in college attendance rates is really being explained by ACT test scores? What if it's only 1%? By contrast, you learn that "number of college fairs attended" explains 80% of the variation (R-squared = .80).This suggests that, despite this official sounding statement, you are doing a fairly poor job of explaining variation in college attendance rates. Only you wouldn't guess that from the statement above. In fact, you might end up focusing lots of attention on ACT prep when in reality you should be focusing on something else--say, college fairs. (Note: This example is hyopthetical!)

Takeaway lesson: When researchers use regression models, ask about the R-squared. Ask how much of the variation they have actually explained. You can also ask about the relative importance of individual variables--e.g., how much did the R-squared increase when you added ACT math scores to your models versus ACT reading?

A warning: Extrapolating beyond the regression line:

In the regression model described above, all the shih tzus are less than a year old. What if a potential adopter wanted to know how much his dog should weigh at age 5? Could she use this model to predict her dog's weight at that age?

The answer is, probably not. Shih tzus grow rapidly during their first few months of life and then they pretty much stop growing unless they eat too many treats and get fat. If your shih tzu continued gaining a pound a month for five years, he would be extremely portly and would probably make the Guinness Book of World records for being the world's biggest shih tzu.

The takeaway lesson here is that regression lines are created based upon the data that we plug into the model. If you try to extend the regression line--i.e. to use this particular regression equation to predict the weight of a dog who is much older than the dogs in your model--you will often run into trouble. This is called **extrapolating beyond the regression line**. You should ask lots of questions if someone is using regression to predict the future or to draw conclusions about a range of data that was not included in the original model (e.g. older shih tzus). Sure, such predictions might be correct and helpful. But those who extrapolate beyond the regression line should make a case for why they believe the pattern summarized by their model will apply to future data or to numbers that are bigger or smaller than those included in their model. As the ads for financial services companies warn, “past performance is no guarantee of future results.”

—

Regression to the Mean/ a.k.a. the regression fallacy:

In most test/re-test situations, the group that does worst on the first test will improve on the second. The group that does best on the first test will do worse on the second. These changes do not signify that one group has really improved and one group has really declined. They are just statistical artifacts.

Why should I care?

Journalists often focus on extremes. *Let's spend time at the worst school. Let's find the kids with the best test scores.* When the worst schools improve and the top kids regress, it is important to account for the regression fallacy rather than automatically assuming that real changes have occurred.

Regression to the mean can also be a problem for accountability programs that sanction or reward schools for year-on-year gains. Purely as a result of regression to the mean, performance at the “worst” schools may improve while performance at the “best” schools declines.

—

Reliability:

The consistency or repeatability of a measurement. If the same perfectly reliable instrument (e.g., test) is used to measure the same thing (e.g., a student's knowledge of 7th grade math) under the same conditions multiple times, identical measurements would result. The more similarity in these multiple measurements, the more reliable the instrument.

—

Standard deviation:

For a set of numbers: A measure of how far the numbers are from the average, i.e., how spread out are the numbers?
Example: Let’s say my test scores in my English class are all pretty similar.

They are 89%, 90%, 92% and 93%.

The standard deviation is pretty small. It is 1.83.

My test scores in math, however, are all over the place.
They are 89%, 65%, 100% and 70%.
The standard deviation of these test scores is much bigger. It is 16.35.

Even if I did not look at the scores on my individual quizzes, I could glance at the standard deviations and gather that my performance is more consistent in English than in math.

Why care?
A standard deviation can be used as a unit to measure how unusual a particular data point is compared to other data points in the set. In the example above, 100% is my most “unusual” math test score. It is more than one standard deviation from the mean of my math test scores, which is 81%. I might want to ask myself what I did differently before I took that test.

A standard deviation can also be a measure of volatility. Consider two schools with the same average test scores. You notice, however that the average scale scores at one of the schools has a much bigger standard deviation, so you decide to snoop around. When you call the principal, you learn that a districtwide gifted and talented program has recently been moved to this school. The high scores of the students in this program are masking the low scores of the rest of the kids.

To calculate the standard deviation in Excel:

1. In the formula box at the top of the spreadsheet, write: =STDEV(A1:Z99) But instead of writing A1,write the cell number of your first piece of data. Instead of writing Z99, write the cell number of your last piece of data.

2. Hit enter. The standard deviation will appear in an empty cell.

Standard error:

The standard error is used to indicate the chances that you would get the same results if you had more data. It is basically a standard deviation [See STANDARD DEVIATION] that accounts for sample size. For instance, let’s say you were trying to figure out the average salary at your news organization. As a reporter, your time and funds are limited so you decide to randomly select ten employees and ask them to tell you their salaries. Much to your dismay, you find the average salary is $10,000. You wonder whether this is the real average salary or if, by chance, you netted some really poorly-paid employees in your random sample. The standard error gives you an idea of how your results might have been different if you had had the time to sample more people. Let's say your standard error is $2,000. That means that the average salary at your newspaper is probably somewhere between $6,000 and $14,000 (with a 95% confidence interval)..

The difference between your real world, limited sample and the ideal situation of multiple samples or large samples is not be due to real differences in the mean salary. The mean salary is a specific figure, even though only the publisher knows what it is and refuses to tell you. The difference that the standard error measures indicates how much your results might change as a function of the people you happened, by chance, to include in each sample.

Warning: A standard error cannot compensate for a biased or non-random sampling process. For instance, if you sampled your friends in the newsroom rather than taking a random sample, the standard error would not help you adjust for the fact that your selection was not random.

SEE ALSO: MEASUREMENT ERROR.

Validity:

It is important to ask about the validity of both tests and studies.

Validity and Testing:

Validity of inferences: Should the judgment made be based on these test results? A test may be perfect but not sufficiently related to the use (e.g., A good eye test may be reliable, but it shouldn't be used to decide whether to award a high school diploma).

Construct validity: Think of this as the overarching validity question, Does the test measure what it's purporting to measure? Construct validity is the degree to which all accumulated evidence supports the intended interpretation of the test scores for the intended purpose.

Validity and Research Studies:

Internal Validity: The causal relationship between the treatment and the outcome. Did in fact the treatment make a difference in this specific instance?

External validity: The ability to generalize a study's results to other settings. Would the results obtained in this instance also be obtained in other similar programs or approaches?