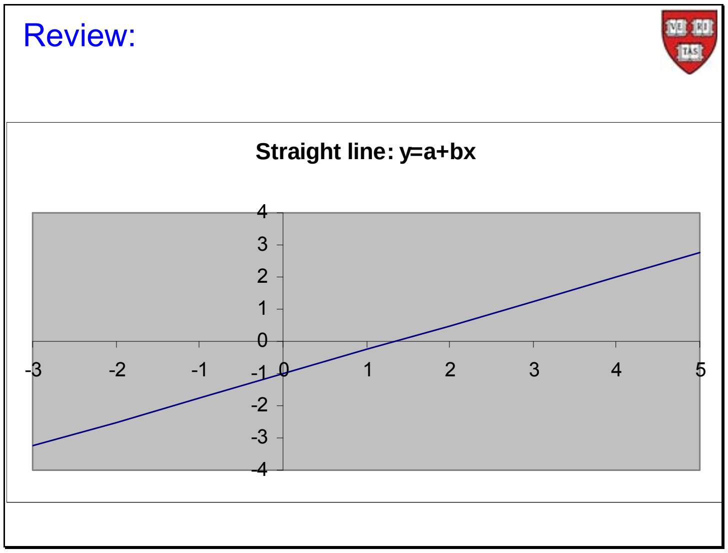
Marcello Pagano

[JOTTER 10 LINEAR REGRESSION]

Simple linear regression, least squares, indicator variables, multiple regression, subset regression



Review:

Today we start on our study of regression; in particular, *linear regression*. It is very close to what we have just done with correlation. It is a continuation of our study of how variables vary together. Now we want to allow more structure and not stop at just two variables.

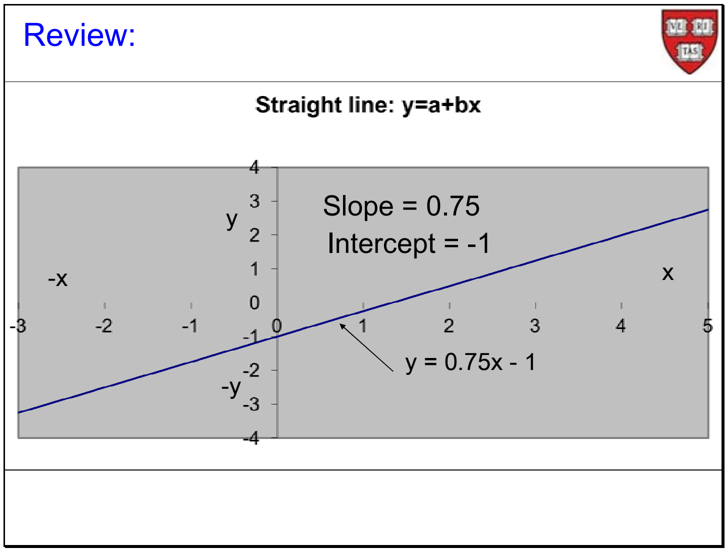
First a very quick review of your high school mathematics, and when you looked at a straight line. Here is a straight line drawn on a Cartesian coordinate system together with its definition. A straight line describes the relationship between two variables, x and y. So given an x I can find a single y, and vice versa, given a y I can find an x. I have a little trouble with this generalization when we have a flat or a vertical line. More about those two extreme cases later, but they are the only two that give us problems.

The reason we are looking at this is that up to now we have been estimating a single population parameter, like the mean, the standard deviation, or the correlation coefficient. Now we are going to define a line in the population and try and estimate it. This is not as complex as it sounds because the line is defined by two parameters, a and b–maybe I should have used the Greek letters α and β to be consistent with our previous work, but I did not want to be fancy. So now instead of estimating a single population parameter, we are being asked to estimate two of them.

The parameter a is the intersect. So, for example, if you put x equals 0 into the equation of the line, you get that y is equal to a. So y=a is where the line crosses the y-axis.

The parameter b is the slope of the line: If we start at a point on the line, and we increase x by one, y gets increased by the amount b.

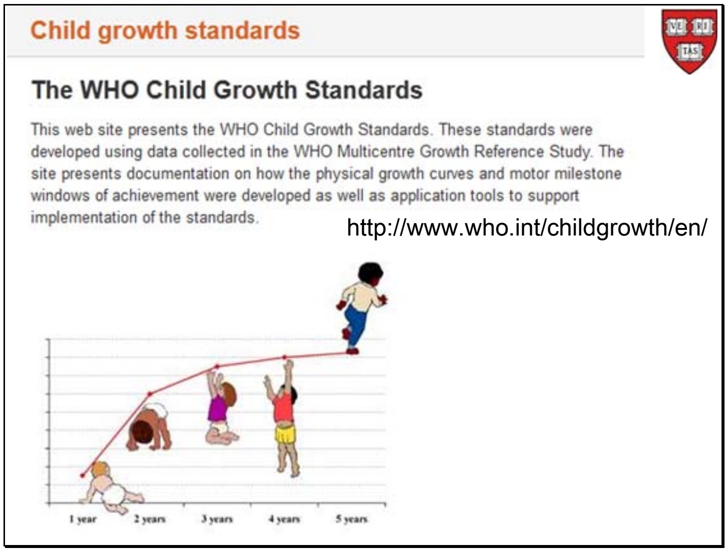
So now I hope all these memories of straight lines have come back to you.



Review:

Straight line: y=a+bx

Let us look at a particular straight line, above, that has slope and intercept as stated.



Child growth standards

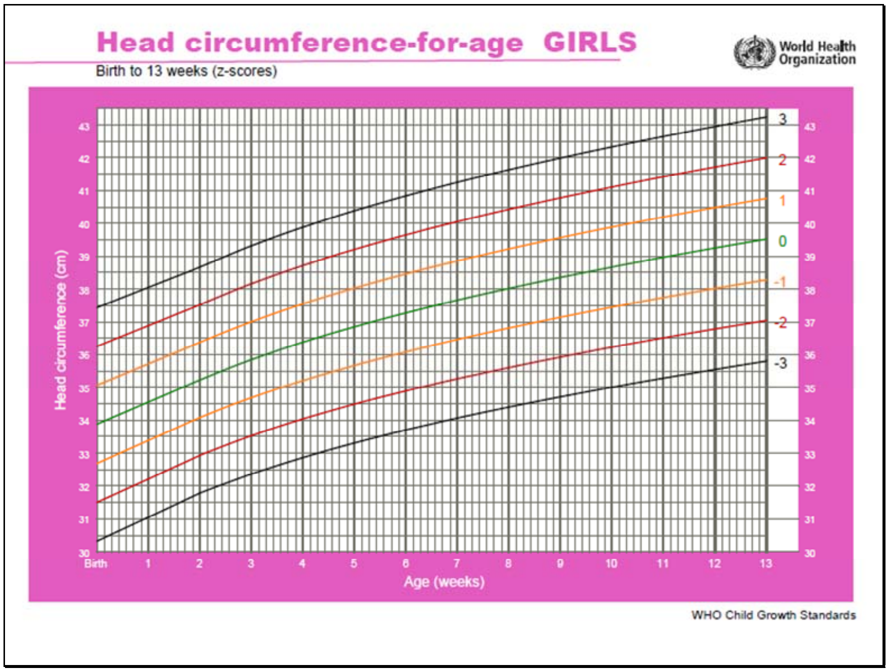
The WHO Child Growth Standards

This web site presents the WHO Child Growth Standards. These standards were developed using data collected in the WHO Multicentre Growth Reference Study. The site presents documentation on how the physical growth curves and motor milestone windows of achievement were developed as well as application tools to support implementation of the standards.

http://www.who.int/childgrowth/en/

The WHO is much concerned about making statements about how we should be growing. Here is a typical picture showing a monotonic relationship between age on the horizontal axis and some generic measure of growth that should be progressing as we get older. We could be talking of weight, head circumference, or any number of properties associated with growth.

These graphs typically show you how things should progress, and since we do not all progress at the same rate, the graphs also show bounds within which "normal" growth is somehow defined. Of course, falling outside these bounds is sometimes used as a warning that something could be amiss.



Here is an example. This is a chart from the WHO website[1] that shows how head circumference increases with age for girls.

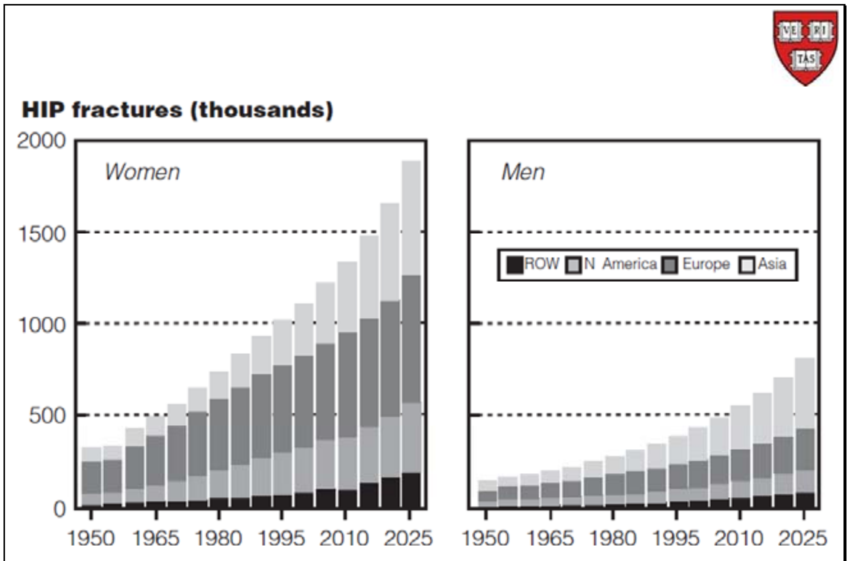
The green line in the middle is the mean. To explain this line: for every fixed age there is a population of girls that age. If we measured the head circumferences for each of these girls and calculated the mean of those numbers, we would get the reading on the green line. So, at birth, for example, that mean is 34 cms, the mean head circumference at birth for baby girls. At eight weeks, the mean is 38 cms, and that is the mean head circumference for baby girls who are eight weeks old. And so on. We get those readings by looking at the scale on that lovely, purply color.

If at each age group we standardize the reading on each girl, get the Z-score—remember, we standardize by subtracting the mean for that age group and then dividing by the standard deviation for that age group—then the green, or mean line is the line when the Z-score is zero. Looking at the right of the chart we see the green line so identified.

We also see the other colored lines identified as Z-score lines for various values of Z (±1 (orange), ±2 (red), and ±3 (blue)). So these lines give us, for each fixed age, predictive intervals, if you believe in the empirical rule, for percentages of the population. In fact, these head circumferences, for each fixed age, have a distribution that is approximately normal. So the orange lines demarcate approximately the middle two-thirds of the population; the red lines demarcate approximately the middle 95% of the population; and the blue lines demarcate approximately the middle 99.8% of the population.

[1] http://www.who.int/childgrowth/standards/second_set/cht_hcfa_girls_z_0_13.pdf

So this chart guides us by showing us how the mean head circumference grows with age, and it also shows us the distributions around these means. It answers the question of what is normal, however we wish to define it, and what is not normal, if we define the latter by measuring the head circumference—for example, we can use it to define macro- and microcephaly.



http://whqlibdoc.who.int/trs/WHO_TRS_921.pdf

This is a common approach of describing maturation. Sometimes a more complex approach is taken. For example, let us focus on osteoporosis.

One troubling manifestation of osteoporosis is bone fractures; for example hip fractures. These graphs come from the WHO to show us that the numbers of hip fractures are increasingly with time around the world.[2]

Be careful, part of the chart is based on real numbers and part of the chart is based on predictions; it goes out to 2025 and here we sit in 2012. Also, it is a problem for both men and women, but it looks like a much bigger problem for women than it is for men, so let us concentrate on women for this discussion.

[2] Note the use of stacked bar charts. Here it works very well, but this graphic technique can sometimes lead to confusing results.

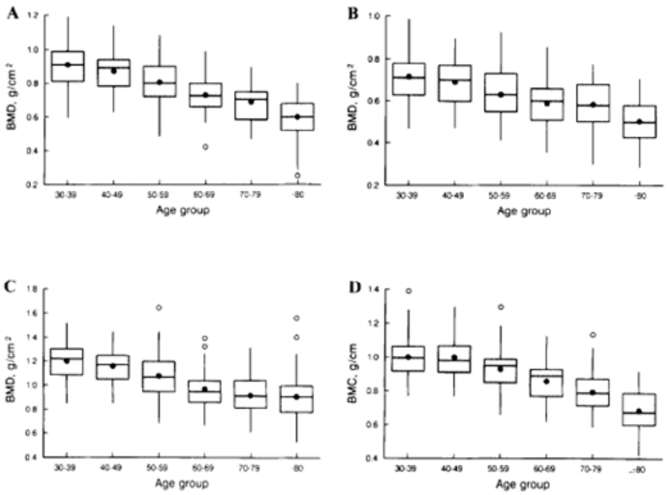
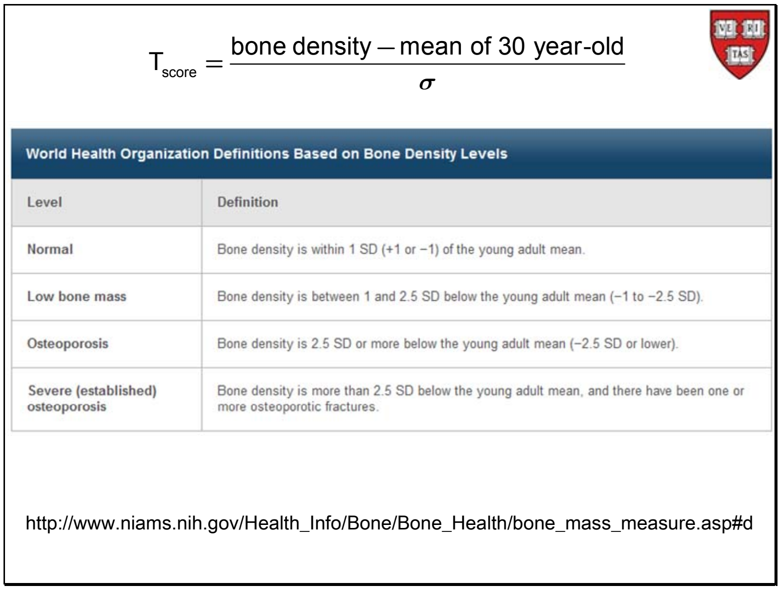


FIG. 1. Distribution of bone mineral density (BMD) at the neck (A) and intertrochanteric (B) regions of the proximal femur and the lumbar spine (C) and bone mineral content (BMC) at the midradius (D) among an age-stratified random sample of Rochester, Minnesota women.[13,31] Boxes extend from the 75th to the 25th percentiles, with a transverse line at the median and a dot at the mean; lines extend to the most extreme values within 1.5 interquartile ranges of the 75th and 25th percentiles; values outside these ranges are indicated individually. The hip and spine measurements were made

Bone density typically decreases with age. Here are four different parts of the body, where this decrease is evident.



$$T_{\text{score}} = \frac{\text{bone density} - \text{mean of 30 year-old}}{\sigma}$$

World Health Organization Definitions Based on Bone Density Levels	
Level	**Definition**
Normal	Bone density is within 1 SD (+1 or −1) of the young adult mean.
Low bone mass	Bone density is between 1 and 2.5 SD below the young adult mean (−1 to −2.5 SD).
Osteoporosis	Bone density is 2.5 SD or more below the young adult mean (−2.5 SD or lower).
Severe (established) osteoporosis	Bone density is more than 2.5 SD below the young adult mean, and there have been one or more osteoporotic fractures.

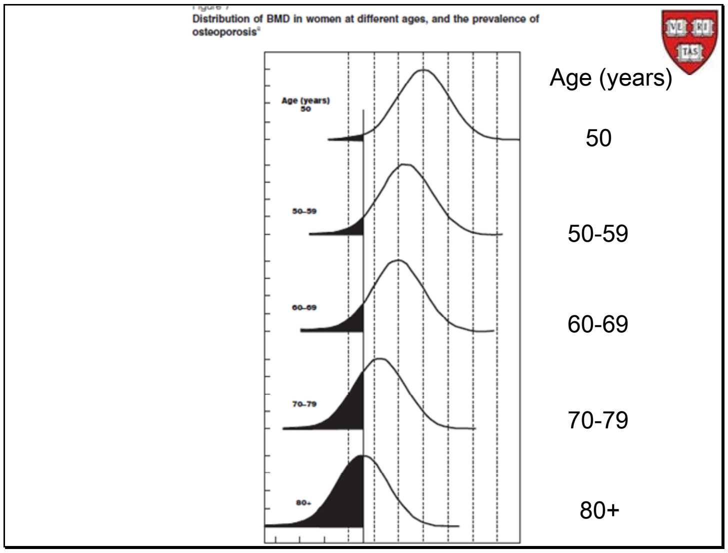
http://www.niams.nih.gov/Health_Info/Bone/Bone_Health/bone_mass_measure.asp#d

What the WHO have done here with the T_{score} is define a way of calculating a measure for each person. This resembles standardization but it is different because they subtract a constant mean, namely a number they get by measuring the mean of a 30 year old. These measures are group specific, so if we are dealing with women, we subtract the mean of 30-year-old *women*.

Had we subtracted the mean of people within the appropriate age group, then we would have the Z-score, just as we did above with head circumference.

The reason for this T-score is to now interpret it as we would a Z-score, but with a difference in the group quantification associated with our favorite numbers; say ±1, ±2, and ±3. Because the mean we use in defining the T-score is not the correct mean to get the 67%, 95%, 99.8% interpretations we are accustomed to with the Z-score, we need other interpretations for the T-score. What we know is that the correct mean goes down with age, so now the T-score is no longer age interpretable without further modification, if we are interested in population percentages.

In practice, the T-score is used for diagnostic purposes: If the T is above -1, then the person is called “Normal”[3]; if the T-score is between -1 and -2.5, then the person is labeled to have “Low bone mass”[4]; and, if the T-score is less than -2.5 then the person is diagnosed to have “Osteoporosis”. The last label is explained above. How many people fall into these categories depends on the age of the person—to repeat, because bone density goes down with age.



Distribution of BMD in women at different ages, and the prevalence of osteoporosis[5]

Bone density is approximately normally distributed with a drift to the left (smaller mean) as age increases, as we see above. The T-score measures distance from the fixed, 30-year-olds mean. So the age-specific proportion of individuals below a fixed point will increase as age increases, as displayed above[5].

With this construct of the T-score and having it define osteoporosis, we see that they are linking bone density with osteoporosis. And they are doing it in such a way that as age increases, you are getting more and more people with osteoporosis, and the speed with which this increase is occurring obeys the law given by the normal distribution and the increase in the black area

[3] The +1 in the slide must be a mistake.

[4] Sometimes labeled Osteopenia.

[5] http://whqlibdoc.who.int/trs/WHO_TRS_921.pdf

above. This is a rather stringent requirement. It dictates how many individuals are to be labeled with normal, osteopenia, osteoporosis, and severe osteoporosis at every age. It would be interesting to see the empirical verification of this classification.

So it is not that they are just trying to scare you with these T-scores, there does seem to be some thought on how these numbers increase. That this speed is correct, I have not been able to find a justification.

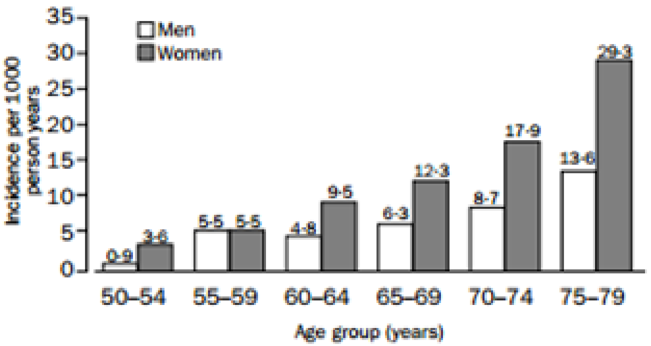


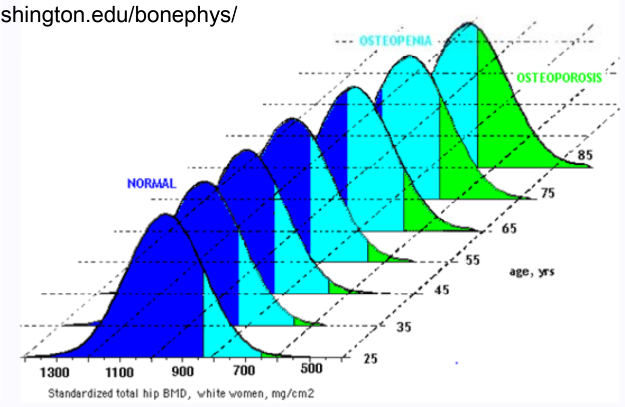
Figure 2: **Incidence of vertebral fractures in women and men from Europe by age group**[48]

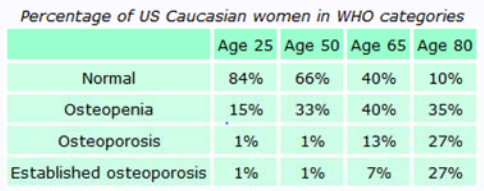
Vertebral fracture was diagnosed from lateral radiographs of the spine and defined as 20% and at least 4 mm reduction in our measure of height in the same vertebral body.[48] Adapted from reference 48 by permission of the Amercian Society for Bone and Mineral Research.

Above are some results in that direction[6].

[6] SR Cummings and LJ Melton, Epidemiology and outcomes of osteoporotic fractures, *The* Lancet, **359•** May 18, 2002 • www.thelancet.com

http://courses.washington.edu/bonephys/opbmd.html





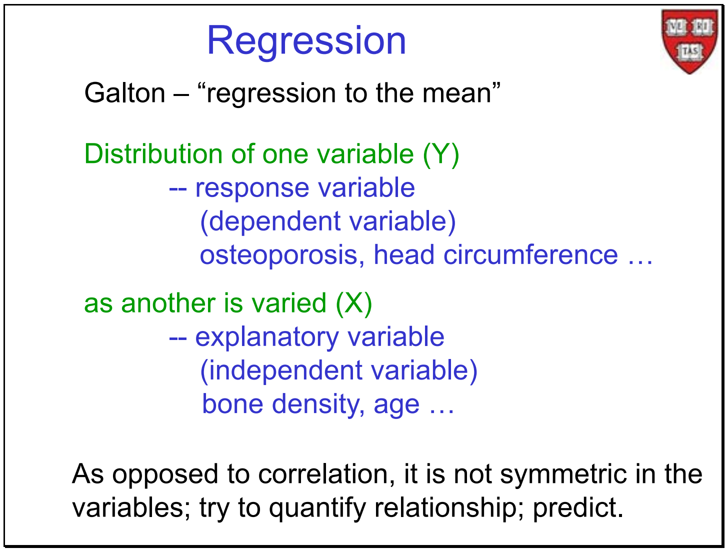
Percentage of US Caucasian women in WHO categories

	Age 25	Age 50	Age 65	Age 80
Normal	84%	66%	40%	10%
Osteopenia	15%	33%	40%	35%
Osteoporosis	1%	1%	13%	27%
Established osteoporosis	1%	1%	7%	27%

Above is a wonderful graphic attempting to show everything we have just been discussing about the T-score.

Note how the distributions are shifting with age, so if the cutoffs remain the same, the proportions in the various classes change as is evidenced by the colors. The distributions change with age. The standard deviations remain the same—homoscedasticity—but the means remain change.

This idea that at every age group we have a distribution is an important one, for the next topic, regression. We have seen this before when we looked at the sex ratio as a function of gestational age; and when we looked at head circumference changing with age. How these changes are linked to each other is our next topic: regression.



Regression

Galton – “regression to the mean”

Distribution of one variable (Y)

-- response variable
(dependent variable)
osteoporosis, head circumference …

as another is varied (X)

-- explanatory variable
(independent variable)
bone density, age …

As opposed to correlation, it is not symmetric in the variables; try to quantify relationship; predict.

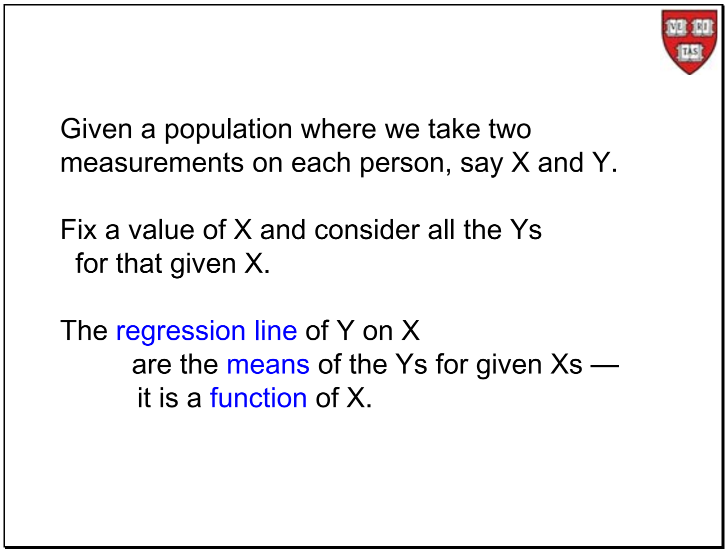
The idea behind regression goes something like this: we look at the distribution of one variable, call it Y—for example, osteoporosis, or head circumference—as another variable, call it X—which might be bone density in the case of osteoporosis, or age in the case of head circumference. We look at the distribution of one of them, Y, to see how it is affected by the other variable X.

So the Y is sometimes called the response, or dependent variable, and X the explanatory, or independent variable. Of interest is how the distributions of the response variable vary as we change our explanatory variable. So, for example, as age increases, how does the distribution of head circumference vary? We start trying to answer this question by first zeroing in on a particular characteristic of the response variable, and that is its mean.

So the question we ask is, how do the means of the Ys vary as the value of X varies? This is the technical meaning of the word regression, and it is due to Francis Galton. He actually originally called it regression to mediocrity, but in those days mediocrity meant mean, so you sometimes see it referred to as regression to the mean.

What he had discovered was that when he measured certain characteristics of children and their reported parents in England, there were interesting relationships. For one, if you looked at sons and reported fathers, the tall man would be associated with tall sons, although they were not as tall as the reported fathers. So too with short sons: the short man would be associated with a short reported son, but the son was not as short as the reported father. So, in both of these cases, when going from one generation to the next, the heights of the younger generation had them becoming “closer to the mean”; a regression to the mean.

Regression is different from correlation, although you will see shortly that they are closely related. Correlation between X and Y is symmetric—it is the same as the correlation between Y and X. Regression is statistically different in that we describe the relation between the mean of the Y as X varies, and that is not the same as describing the mean of the X as Y varies. Also, we use regression differently in that the roles of Y and X are different; X is more the independent variable, the explanatory variable. It is the variable we may be able to set or control some, whereas Y is what results from having set the X. Y is the outcome or dependent variable.

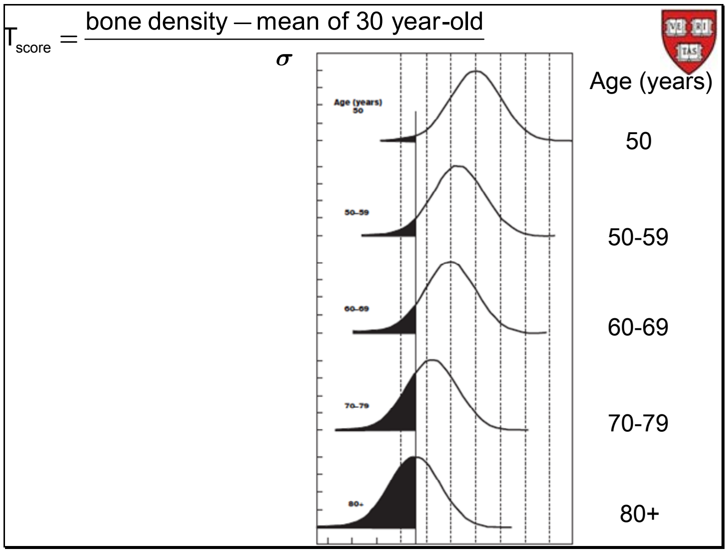


Given a population where we take two measurements on each person, say X and Y.

Fix a value of X and consider all the Ys for that given X.

The regression line of Y on X are the means of the Ys for given Xs — it is a function of X.

Here is a formal definition of the regression line. We usually abbreviate it to regression, but we really are referring to a regression line.

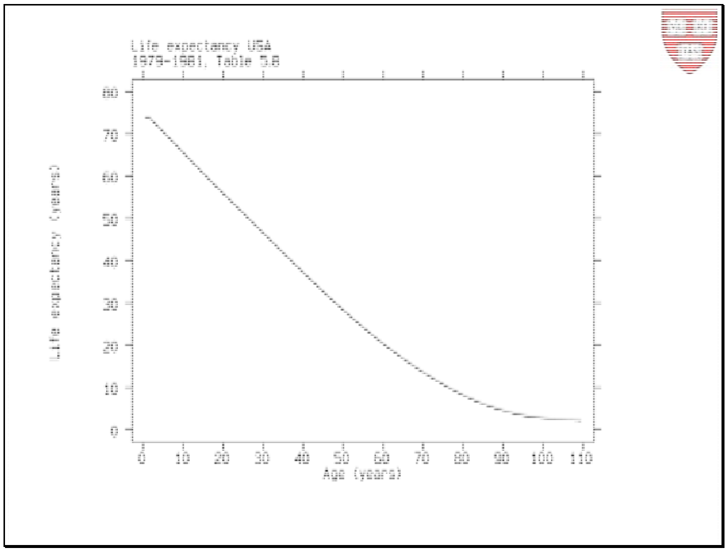


$$T_{score} = \frac{\text{bone density} - \text{mean of 30 year-old}}{\sigma}$$

[7]

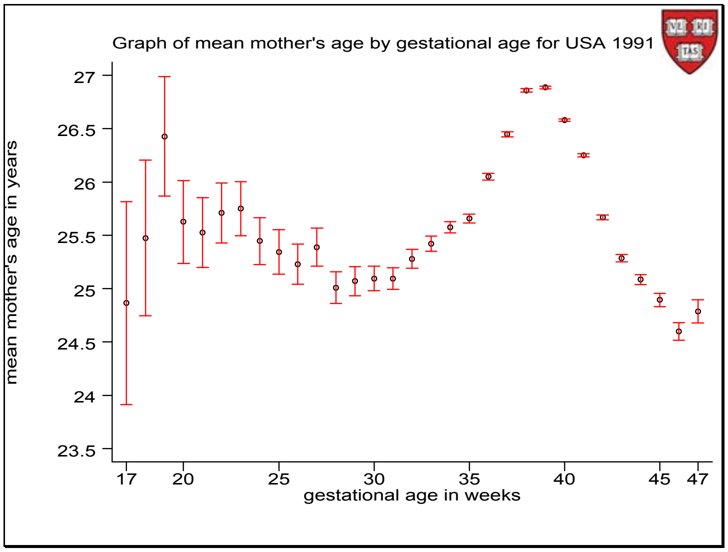
[7] http://whqlibdoc.who.int/trs/WHO_TRS_921.pdf

We can get an idea of what the regression line is from the above. This data set is less than ideal to depict the thought, because the age is not shown exactly, but rather in groups. Take the pedagogical license of acting as if the age was at approximately the center of the intervals: 50, 55, 65, 75 and 85. Now visualize a graph with those ages plotted on the horizontal axis, and on the vertical axis the means of these normal density functions (their centers). Join up those points. That is your regression line of bone density on age.



Here is another example. Look at this curve where for each age on the horizontal, the vertical value of the curve gives the mean life remaining for persons that age. That is the regression line of residual survival on age—how much time, on average, a person still has to live.

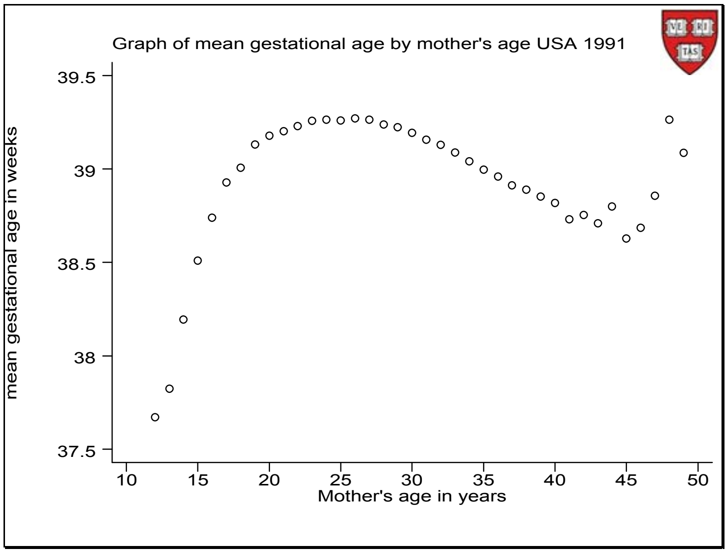
Note that the line is quite straight from about age two to about age sixty. If that were so then we would say we have linear regression over that region. We are going to be focusing on linear regression at first.



Here is another example of a regression line (obtainable by joining the dots.) This is the same group for which we earlier saw the sex ratio at birth—the group of about 4.2 million singleton births in one year, 1991, in the US. This time we look at the mean of the mothers' ages for each gestational age.

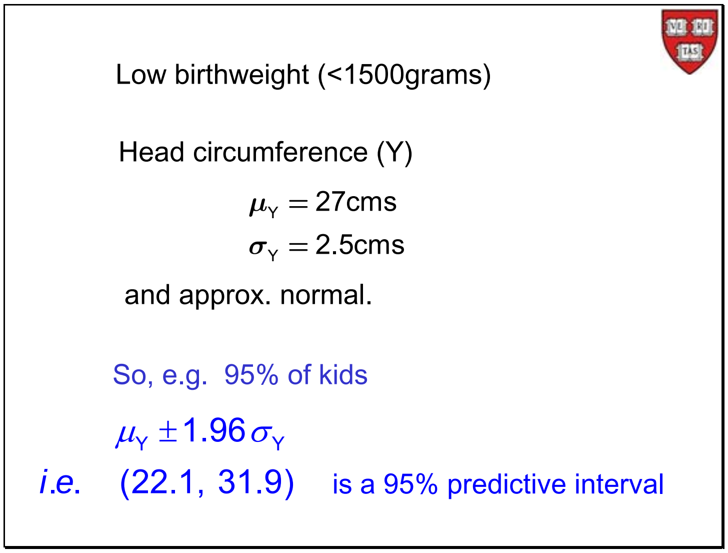
We also see the plus and minus two standard errors around the mean ages to see the 95% confidence intervals for the means. The widths of these intervals are quite tiny because the sample sizes are so large.

Given the prior statement about what is the meaning of independent and dependent variable, one could argue that we have the classification in reverse.



When we do interchange the roles, this is the regression line of gestational age of the baby on the mothers' ages. The curve is quite smooth until we reach mothers' ages in the mid-forties, but these data reflect conditions in the early nineties, so the points on the right hand side are means of very few mothers.

The whole regression is not linear, but one can see two or three linear segments in this graph. We focus on linear regression for the rest of this chapter.



Low birthweight (<1500grams)

Head circumference (Y)

$\mu_Y = 27$cms

$\sigma_Y = 2.5$cms

and approx. normal.

So, e.g. 95% of kids

$\mu_Y \pm 1.96\,\sigma_Y$

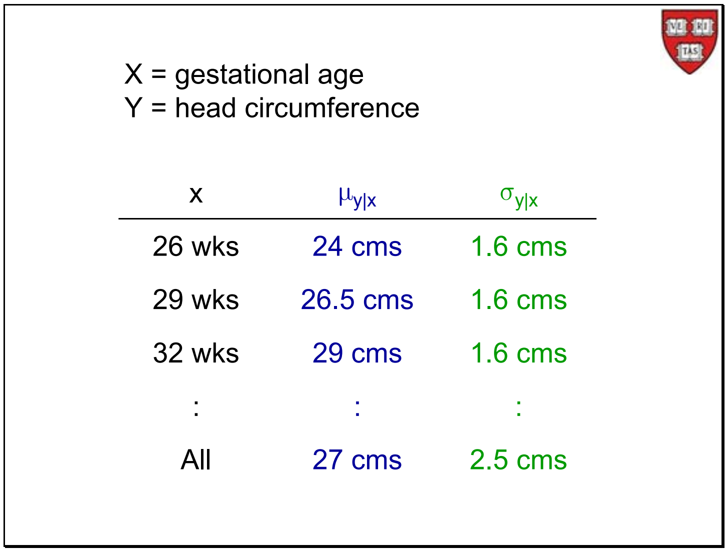
i.e. (22.1, 31.9) is a 95% predictive interval

First we look at what we call *simple linear regression*. It is so called because we have a single outcome variable, a single explanatory variable, and the regression line is straight. We extend this shortly to the situation when we have more than one explanatory variable, and then we consider a certain class of non-linear regressions.

Return now to the group of low birth weight infants. Here we assume that they weighed less than 1,500 grams at birth. Let us look at the distributions of their head circumferences, and how these distributions vary with gestational age. So it is a little similar to the data from the WHO, we looked at above, except that this is a breakdown of a certain subset of that population, at birth.

What we find is that this population is approximately normally distributed with mean 27 centimeters and a standard deviation of 2.5 centimeters. So we can set up predictive intervals etcetera. For example, the 95% predictive interval would be roughly from 22.1 to 31.9 centimeters.

The question we can now ask is can we be more precise in our predictive interval? Is there some other information we can bring to bear that will allow us to construct a tighter predictive interval, let us say? In other words, can we legitimately use a smaller standard deviation? This is what regression allows us to accomplish, and the answer is yes, if we can find a good explanatory variable.



X = gestational age
Y = head circumference

x	$\mu_{y\|x}$	$\sigma_{y\|x}$
26 wks	24 cms	1.6 cms
29 wks	26.5 cms	1.6 cms
32 wks	29 cms	1.6 cms
:	:	:
All	27 cms	2.5 cms

If we concentrate on babies born at 26 weeks gestational age, we find that their head circumferences are approximately normally distributed with mean 24 centimeters and a standard deviation around that of 1.6 centimeters. So immediately we see that these kiddies have smaller heads (mean 24 versus 27), and more homogeneous (standard deviation 1.6 versus 2.5) than when we ignore their gestational age, and just consider all these kids as a single group. So it looks like gestational age is going to earn the privilege of being called an explanatory variable, in this setting.

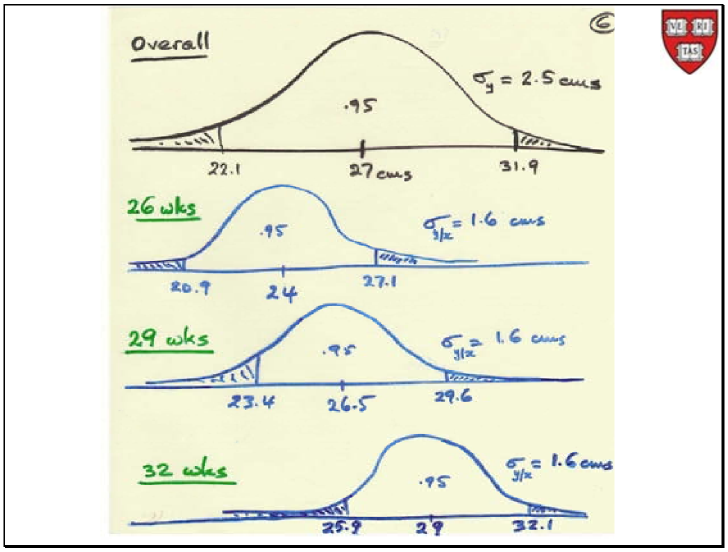
Before we continue, note the new notation we introduce. The mean and standard deviations—μ and σ, Greek letters because we are talking about the population of these kids for the moment—have subscripts to show that we are talking of the head circumferences, our ys, and their dependencies on the gestational age, x.

Now concentrate on the 29-weekers. Their mean head circumference is slightly larger than the 26-weekers—26.5 centimeters versus 24 centimeters—although the standard deviation remains constant at 1.6 centimeters.

Similarly with the 32-weekers, the mean head circumference edges up to 29 centimeters and the standard deviation remains at 1.6 centimeters.

So overall it seems that the mean head circumference, the regression line, goes up with gestational age, and we have homoscedasticity (constant standard deviations) around this regression line. This homoscedasticity is a special property that does not always hold.

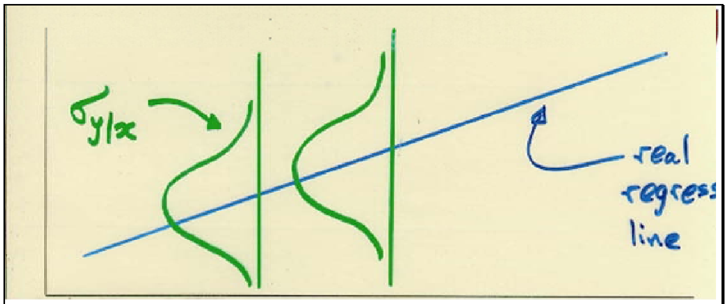
Suppose this is an accurate representation of the whole; namely, this regression line and homoscedasticity holds for the other gestational ages. Then by taking the gestational age of the baby into account we have a more homogeneous group and get a tighter predictive interval for the head circumferences.



Schematically, here is what is happening. This is my attempt at free hand. Overall, when we consider the head circumference of all low birth weight infants, the appropriate distribution is the black one on top; a normal with mean 27 centimeters and standard deviation 2.5 centimeters. Breaking this population down into groups defined by the explanatory variable, the gestational age of the infant at birth, we get the blue distributions displayed above; each has standard deviation 1.6 centimeters (homoscedasticity) that is smaller than the overall standard deviation of 2.5 centimeters, and thus the groups are more homogeneous, but their means vary. These distributions drift from the left of the picture to the right as the gestational age increases to reflect the fact that the head circumferences increase with gestational age.

This drift means that the predictive intervals will also drift from left to right. Of course there are kids at 27 weeks, 28 weeks, etcetera but my drawing skills only extend so far and had I tried to include them all in the picture we would have ended up with a mess[8].

[8] or a Jackson Pollock masterpiece?



If X and Y are normally distributed

$$\sigma^2_{y|x} = (1 - \rho^2)\,\sigma^2_y$$

where ρ is the correlation coefficient between X and Y.

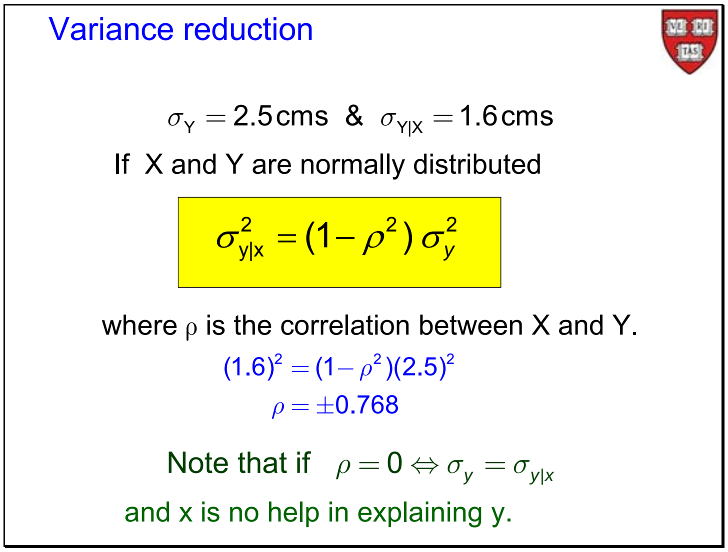
Now focus on the means of these groups and plot them against gestational age. That is the regression line of head circumferences on gestational age. The means fall on a straight line. We return to that below, but first concentrate on the standard deviations of these various groups. What is the relation between these conditional standard deviations and the overall standard deviation?

If we have linear regression and the distributions are normal, as we do here, and we have homoscedasticity, as we do here, then the variance of the gestational-age-specific distributions are related to the overall variance (ignoring gestational age, just considering all these babies together) by the formula above. Thus the correlation coefficient squared reflects the proportional reduction in the variance afforded by considering the explanatory variable, gestational age, in the argument.

A special case of this is when the correlation is plus or minus one, in which case this, so called, *residual variance,* is zero. (This forms the basis of the tip we used to guess the values of the correlation coefficients when faced with those four scatter plots in the correlation game!)

The other special value of the correlation coefficient is when it is zero. Then there is no reduction in the variance. That means that the independent, or explanatory, variable in question is uncorrelated with the outcome variable and linear regression would not be helpful in defining more homogeneous sub-groups. The explanatory variable does not explain.

So here is connection number one between the correlation coefficient and regression: It is the relative reduction in the variance—or standard deviation—of your measurement, when you introduce the explanatory variable into the argument. When we consider the estimation of the regression line, below, you will be introduced to a statistic, R-squared, that estimates the squared correlation coefficient above. Its importance is predicated by the above equation.



Variance reduction

$$\sigma_Y = 2.5\,\text{cms} \;\;\&\;\; \sigma_{Y|X} = 1.6\,\text{cms}$$

If X and Y are normally distributed

$$\boxed{\sigma^2_{y|x} = (1-\rho^2)\,\sigma^2_y}$$

where ρ is the correlation between X and Y.

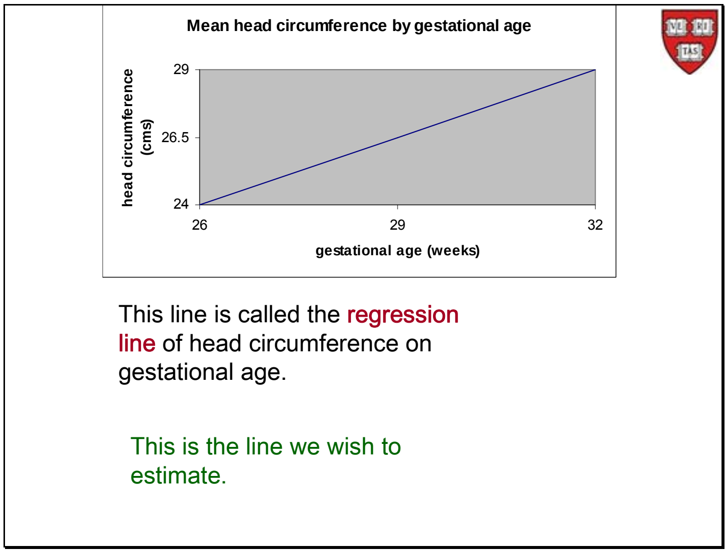
$$(1.6)^2 = (1-\rho^2)(2.5)^2$$

$$\rho = \pm 0.768$$

Note that if $\rho = 0 \Leftrightarrow \sigma_y = \sigma_{y|x}$

and x is no help in explaining y.

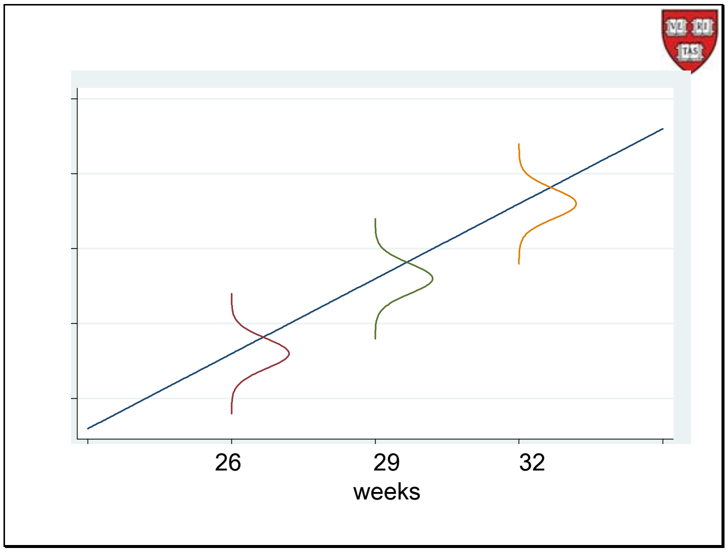
We can use the equation above to calculate the value of the correlation squared. To then determine which square root to use, we determine the sign by arguing that the head circumference increases with gestational age and thus the correlation is positive.



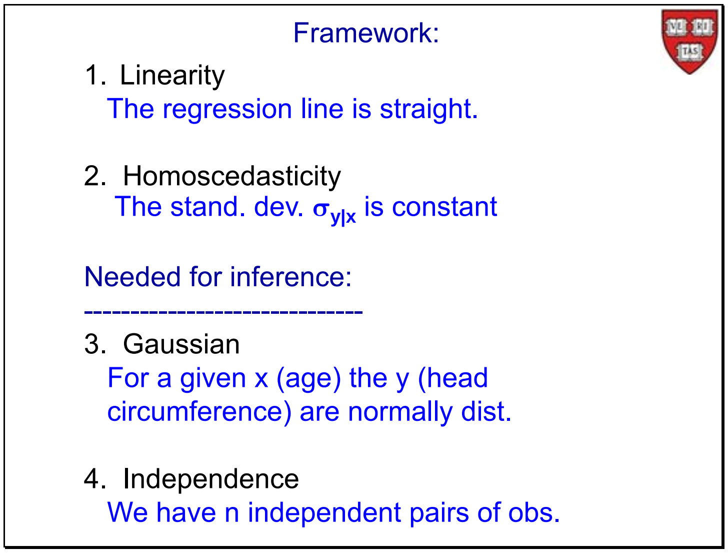
This line is called the regression line of head circumference on gestational age.

This is the line we wish to estimate.

Returning to the regression line of head circumference on gestational age, this is the line that we eventually want to estimate. Up to now we have inferred values of single parameters such as a mean, or prevalence, or variance, or an odds-ratio, a correlation, etcetera. Now we wish to estimate a relationship in the population that is represented by a line. So inference becomes a little more complex.



The line we want is the line where the means of the distributions lie. To depict that we have coming out of the paper, in a three dimensional plot that not great, but there it is. Actually, I have to thank Nick Cox for drawing this for me.



Framework:

1. Linearity
   The regression line is straight.

2. Homoscedasticity
   The stand. dev. $\sigma_{y|x}$ is constant

Needed for inference:

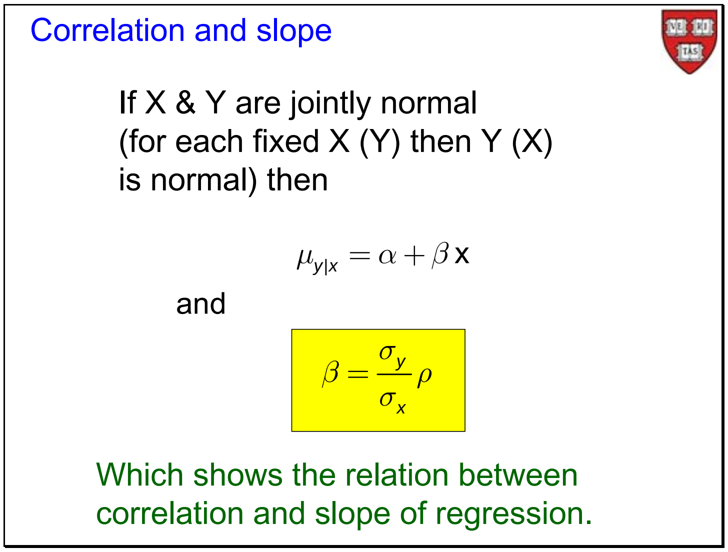
----

3. Gaussian
   For a given x (age) the y (head circumference) are normally dist.

4. Independence
   We have n independent pairs of obs.

The framework within which we operate is: (i) linearity, so we assume that the regression line we want is a straight line; (ii) homoscedasticity, so we assume that each of the distributions coming out of the paper, above, has the same standard deviation. These two assumptions are sufficient to get us going and estimate the regression line in a reasonable fashion. Subsequently to make further inference, we assume that (iii) we have normal data, i.e. the distributions coming out of the paper, above, are each normal; and (iv) our usual assumption that we have independent data.

All of these assumptions can be relaxed, but we delay that to your next course on regression.



Correlation and slope

If X & Y are jointly normal
(for each fixed X (Y) then Y (X)
is normal) then

$$\mu_{y|x} = \alpha + \beta x$$

and

$$\boxed{\beta = \frac{\sigma_y}{\sigma_x}\rho}$$

Which shows the relation between
correlation and slope of regression.

Now we come to connection number two between the correlation coefficient and regression: If we look at our regression line and call the slope β, then there we have the relation between the slope of the regression line and the correlation coefficient. So if Y and X have the same standard deviation—for example if we have standardized both variables so they each now have standard deviation equal to one—then, in that particular case, the correlation coefficient is the slope of the regression line.

In general, the interesting relationship is when the correlation coefficient is zero (uncorrelated) then we see that the slope of the regression line is zero. So the population means of the Ys does not depend on the explanatory variable X. [Theory alert: skip to next paragraph if not interested in theory.] With the conditions we have imposed here, namely normality and homoscedasticity, then we see, since the whole distribution in the normal case is determined by the mean and the variance, that flat regression line, or uncorrelated is synonymous with independence between X and Y. [End of theory alert.]

Equivalence

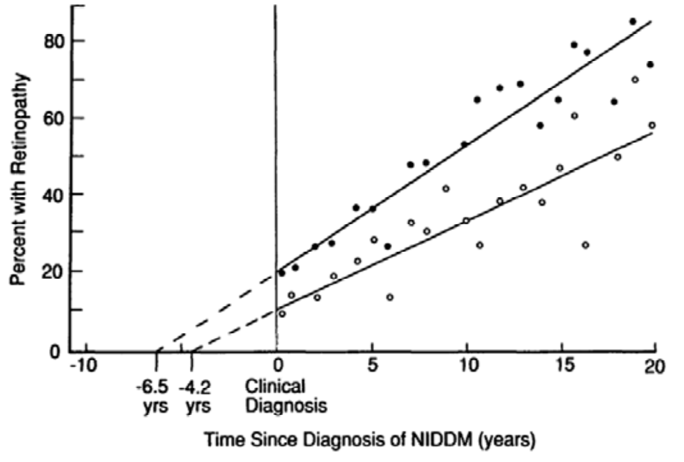
So with normal data the following 3 hypotheses are equivalent:

$$H_0: \quad \rho = 0$$
$$\Leftrightarrow$$
$$H_0: \quad \beta = 0$$
$$\Leftrightarrow$$
$$H_0: \quad \sigma_y = \sigma_{y|x}$$

Collecting those two relations between the correlation coefficient and the regression line together, we come up with the equivalence of these three null hypotheses.

Least Squares

Onset of NIDDM Occurs at Least 4-7 Yrs Before Clinical Diagnosis



MI HARRIS R KLEIN TA WELBORN MW KNUIMAN DIABETES CARE **15**, 7, JULY 1992

All too often in science we are faced with the generic, here are some points through which I would like to draw a straight line, problem and that sets us off in search of the line that best represents, or summarizes, those points. What we mean by that word `best' sounds, of course, subjective.

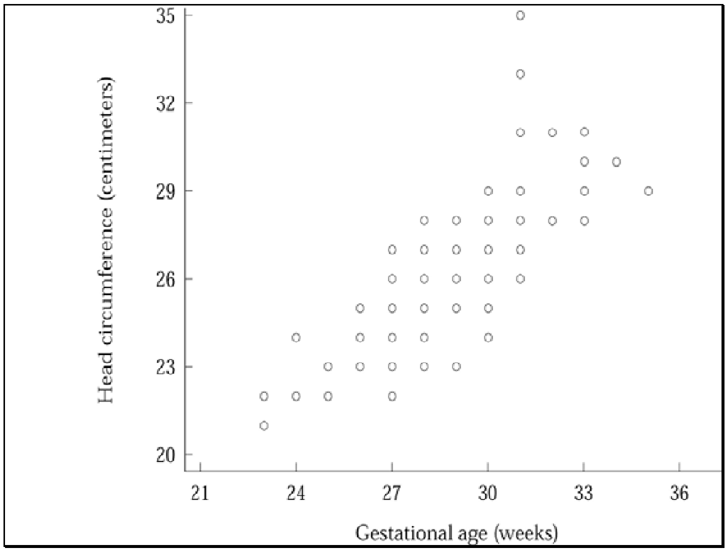
In regression, we seek to determine the regression line associated between an outcome variable and an explanatory variable. Frequently, we cannot afford to fix the value of the explanatory x and find a huge number of ys to determine their mean. Then fix another, different x and repeat the process, and so on and then lay all the means out only to discover that because of sampling variability, experimental error, what have you, the points still do not lie on a perfect straight line. So how do we obtain this population regression line?

We are going to proceed with how we have learnt to do inference. We are going to take a sample of points and infer from them what the population parameters are. In this case the population parameter is a straight line defined by two parameters, the intercept and the slope of the line.

So let us return to the generic curve fitting problem where we have points through which we would like to fit a straight line. Above we see an example where they looked at the prevalence of people with diabetes related retinopathy plotted as a function of time since diagnosis with Type II diabetes mellitus (NIDDM)[9].

This study was done in two countries—in Australia and in the US—and we see two sets of points, one set for each country. Their argument is that if we fit a straight line through those points, we would be capturing the relationship that seems to exist when considering the increase in prevalence with time. They then proceed to extend the line to the left of the zero, the time of clinical diagnosis of NIDDM, to argue that Type II diabetes exists before current clinical diagnoses are done—about 4.2 years in the one and 6.5 years in the other country—as seen by this incubation period for retinopathy. We leave that argument to the experts, and concentrate on the line fitting issue. So we concentrate on the generic problem of trying to pass a straight line through a collection of points.

[9] MI Harris, R Klein, TA Welborn, MW Knuiman, Onset of NIDDM Occurs at Least 4-7 Yr Before Clinical Diagnosis**,** *DIABETES CARE*, **15**, # 7, JULY 1992



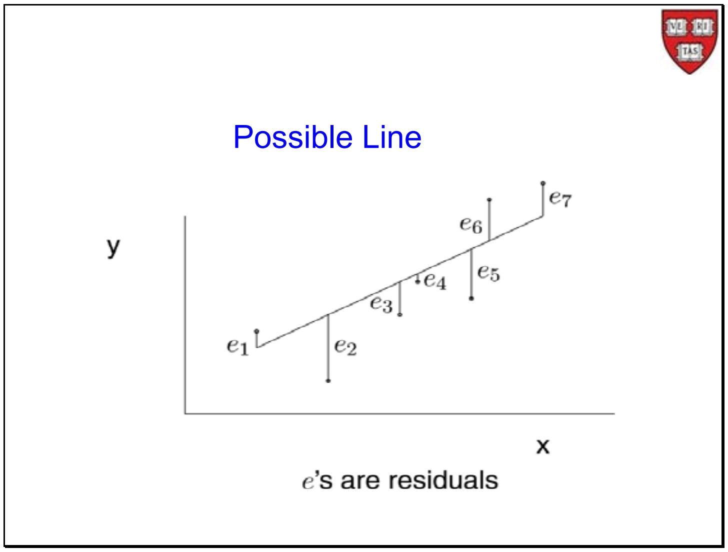
So here are head circumferences for 100 low birth weight, premature babies we introduced above. Care should be taken in reading this graph because some points represent more than one baby. See Page 27, below, for a jittered version of this graph.

For each gestational age we could find the mean head circumferences and fit a straight line through those points, but chances are they would not lie on a straight line. So let us forego that approach and just treat these as general points and see where that leads us.

There are an infinite number of lines that we could draw and claim for each one that it "represent" these points. Heuristically, on the basis of an eyeball test, sometimes facetiously called the intraocular test if you want to impress someone, we can discard a large number of these lines. In an attempt to devise a more objective method for choosing between candidate lines, for each line we draw we can see how well it does by each point by calculating the vertical distance between the line and the point. Ideally we would love all those distances to be zero, but that is not going to happen, unless the points all lie on a straight line.

The length of these vertical distances between each point and the straight line are called *residuals*. Since we cannot make them all zero, we can seek to make them as small as possible, and since we are talking about a number of them, we can consider their sum. If we do that we run into the very same problem we ran into when we looked at the variance. Then we had that when we subtracted the mean from every observation then the sum became zero, so too here, any line that goes through the center of gravity—the point $(\bar{x}, \bar{y})$—results in the sum of the residuals being zero, because the positive residuals cancel out the negative residuals. (What you consider a positive and what you consider a negative residual depends on whether you are calculating distance from the line to the point, or the distance from the point to the line, and whether up is plus and down minus, or *vice versa*.)

So get rid of the signs of the residuals!



Possible Line

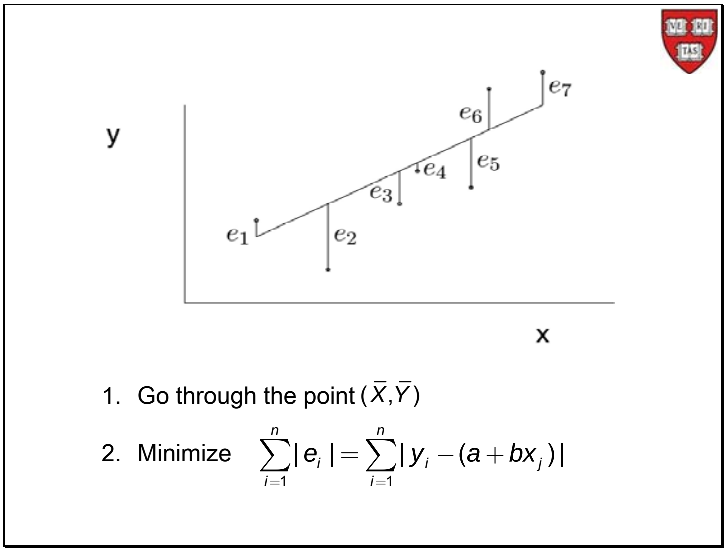
e's are residuals

So here's, for example, e_1 through e_7. Look at the absolute value of those residuals and make their sum as small as you can.

Ruggero Giuseppe Boscovich 1711-1787

And that was the brilliant idea of Ruggero Giuseppe Boscovich[10]. He was attempting to estimate the distance to the moon, and he had a number of conflicting measures, when he came up with this idea.

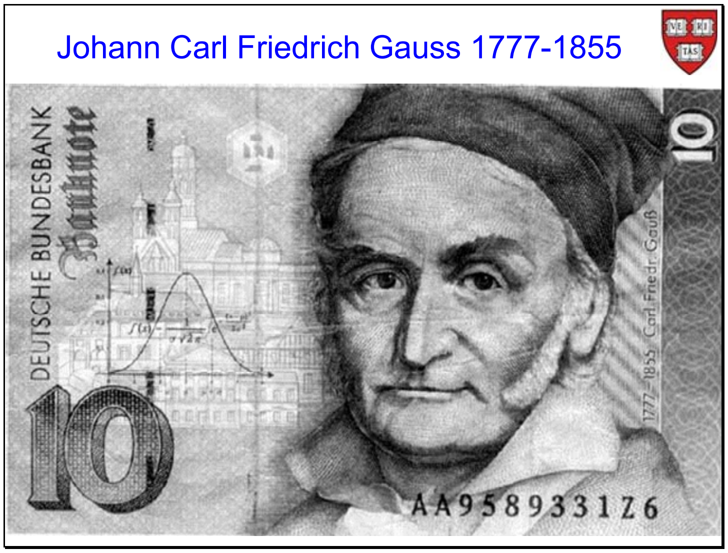
[10] The article by Churchill Eisenhart, "Boscovich and the combination of observations" in *Roger Joseph Boscovich, S.J., F.R.S., 1711-1787 : Studies of His Life and Work on the 250th Anniversary of His Birth*, Whyte, Lancelot Law, Ed, Fordham University Press., New York, 1961.



1. Go through the point $(\bar{X},\bar{Y})$
2. Minimize $\sum_{i=1}^{n}|e_i| = \sum_{i=1}^{n}|y_i-(a+bx_j)|$

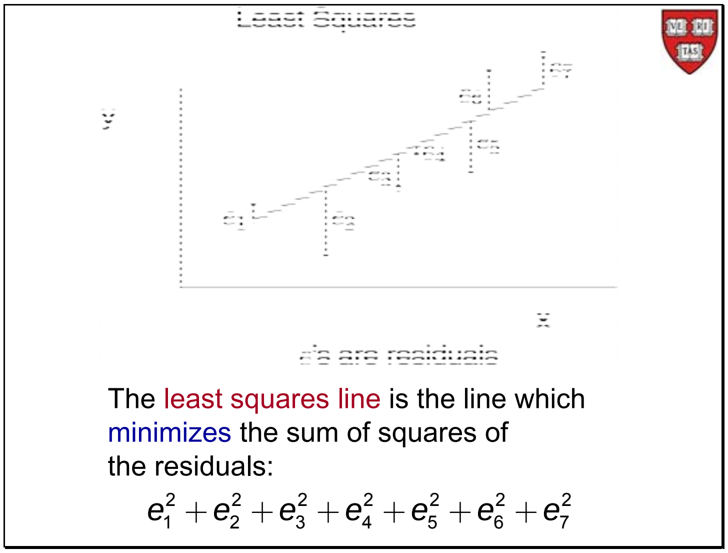
Not only did Boscovich have the idea, but he also gave the algorithm for actually calculating the best straight line; *viz*. the line that goes through the center of gravity and minimizes the sum of the absolute values of the residuals.

There was nothing wrong with his solution, it is truly a lovely algorithm and the basis is very strong, but it did not extend to more than one explanatory variable. As a result it was superseded by the idea of Gauss—although, as usual, there is a dispute about who actually invented it![11]


Johann Carl Friedrich Gauss 1777-1855

[11] Stephen M. Stigler, Gauss and the Invention of Least Squares, *Annals Statistics*, **9**, 1981, 465-474.

Remember Gauss? We looked at this Deutsche Mark because of the normal curve.



Least Squares

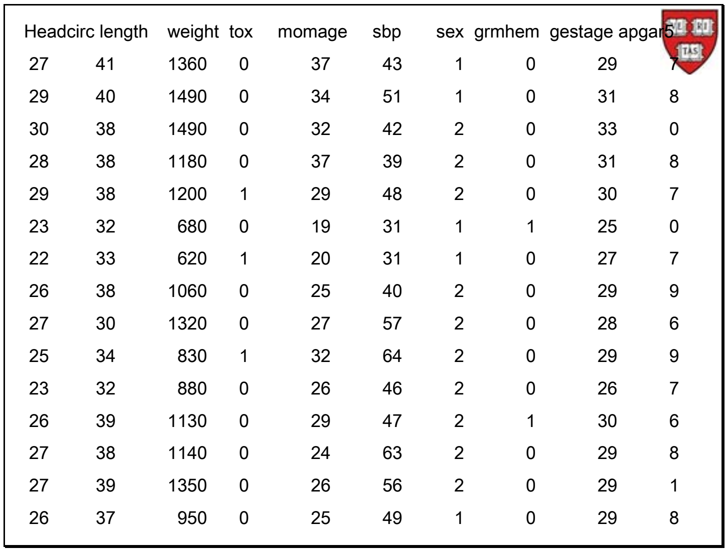
e's are residuals

The least squares line is the line which
minimizes the sum of squares of
the residuals:

$$e_1^2 + e_2^2 + e_3^2 + e_4^2 + e_5^2 + e_6^2 + e_7^2$$

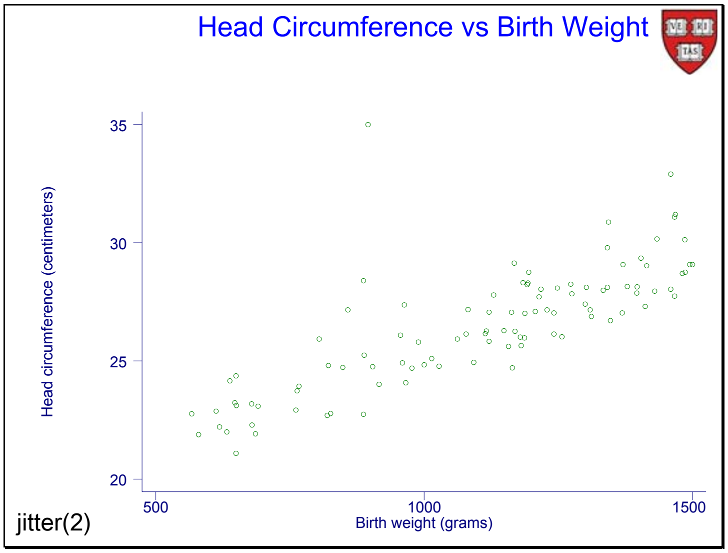
Gauss not only suggested that we minimize the sum of the squares of the residuals—just like when we looked at the variance, we squared the distances from the mean before averaging them out to get rid of the problem of the positive residuals cancelling out the negative ones—but he also provided a wonderfully easy way to actually calculate the slope and intercept of such a line—which, of course, is what Stata does for us. His line, just as Boscovitch's, also goes through the center of gravity.

So amongst all lines that go through the center of gravity, Boscovitch's line minimizes the sum of the absolute residuals, and Gauss's least squares line minimizes the sum of the squares of the residuals. This is a little bit of a shortcoming of least squares in that large residuals get their values squared, thus making them even larger, and this allows them to exert too much influence sometimes. More about that shortly when you explore least squares, below.



Headcirc	length	weight	tox	momage	sbp	sex	grmhem	gestage	apgar5
27	41	1360	0	37	43	1	0	29	7
29	40	1490	0	34	51	1	0	31	8
30	38	1490	0	32	42	2	0	33	0
28	38	1180	0	37	39	2	0	31	8
29	38	1200	1	29	48	2	0	30	7
23	32	680	0	19	31	1	1	25	0
22	33	620	1	20	31	1	0	27	7
26	38	1060	0	25	40	2	0	29	9
27	30	1320	0	27	57	2	0	28	6
25	34	830	1	32	64	2	0	29	9
23	32	880	0	26	46	2	0	26	7
26	39	1130	0	29	47	2	1	30	6
27	38	1140	0	24	63	2	0	29	8
27	39	1350	0	26	56	2	0	29	1
26	37	950	0	25	49	1	0	29	8

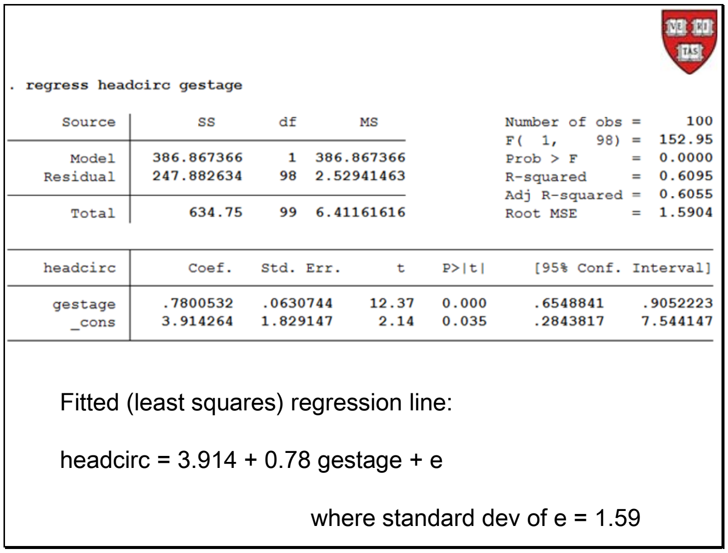
Here are the first fifteen observations in the dataset, lowbwt (low birth weight) a sample of 100 babies whose birth weight is less than 1500 grams. The first variable (headcirc) is head circumference at birth. The next (length) is the length (height, except babies do not stand up!) of the baby in cms, followed by the babies' birth weight (weight) in grams. Whether the mother was toxemic (1) or not (0) is noted in the next variable (tox). Her age is recorded in the next variable (momage) as is her systolic blood pressure (sbp) in mm/Hg. The babies sex is next recorded (sex) where 0=female. Next is recorded whether the baby had a brain hemorrhage, in the germinal matrix (grmhem) with 1=yes. The baby's gestational age at birth is reported in weeks (gestage), and finally the Apgar score at 5 minutes is in the last variable (apgar5).



Head Circumference vs Birth Weight

So here are the plotted data, except that the data has been jittered. I refer you to the Stata manual to see what that means exactly, but roughly what that means is that each data point is moved a smidgen so that no two points fall exactly on each other, since if that were the case it would look like we had a much smaller data set, as happened on Page 23.

The relationship between head circumference and birth weight is arguably linear in this scatter plot.



```
. regress headcirc gestage

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  152.95
       Model |  386.867366     1  386.867366           Prob > F      =  0.0000
    Residual |  247.882634    98  2.52941463           R-squared     =  0.6095
-------------+------------------------------           Adj R-squared =  0.6055
       Total |      634.75    99  6.41161616           Root MSE      =  1.5904

------------------------------------------------------------------------------
    headcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gestage |   .7800532   .0630744    12.37   0.000     .6548841    .9052223
       _cons |   3.914264   1.829147     2.14   0.035     .2843817    7.544147
------------------------------------------------------------------------------
```

Fitted (least squares) regression line:

headcirc = 3.914 + 0.78 gestage + e

where standard dev of e = 1.59

The Stata command for fitting the least squares line is simply *regress* followed first by the response variable (*headcirc* in this case) and then by the explanatory variable (*gestage* in this case). What Stata then does is it calculates the least squares line for us, and provides us with the statistics to allow us to go ahead with the inference we wish to make.

From this we first see that there are 100 data points. Then the bottom two lines tell us about the two variables defining the regression line estimate. The bottom line is the constant (*_cons*) and the line above that is for the gestational age (*gestage*) coefficient.

So we see that the least squares line has constant 3.914 and slope 0.78. That means that for every week extra gestational age, the head circumference at birth will increase by 0.78 centimeters.

When we think of these two numbers as estimates of population parameters, we can see their standard errors, their t-statistic to test the null that each is zero, in turn, and the associated p-values (0.000 for the coefficient associated with gestational age, and 0.035 for the constant). We are usually not concerned with the constant term in situations like this, since it is telling us about something, crossing the axis, at gestational age zero which is meaningless and far outside the experimental region, anyway. The 95% confidence interval for the gestational age coefficient in the population regression line is (0.65, 0.91). So we would reject the null hypothesis that the head circumference is not impacted by gestational age, on the basis of these 100 observations.

The last statistic to look at is the R-squared, as advertised above, which is the estimate of the correlation coefficient squared, or reduction in variance. I would recommend looking at the adjusted R-squared, especially when we have more than one explanatory variable.

___

I leave the transcript of what happened with the visit to the National Council of Teachers of Mathematics website, but strongly urge you to watch that video and visit the site.

___

So let's go to a website run by the National Council of Teachers of Mathematics. And they've got this lovely little applet here that we can use to get some idea of how least squares works. In particular, what I'd like to show you is that least squares can be very sensitive to 1 or 2 points. And since we saw that the least squares line is also intimately related with the correlation coefficient, this will reinforce the statement that I made last week that the correlation coefficient is also sensitive to outliers.

So here's what we do. We go and we plonk down some points. So here are a few points. Or so here there's, what, two, four, five. There's another point. And then you can say, show the line. So this line is the least squares line through those five points. And we see that it's almost a perfect fit. R is equal to 1 with an intercept of 4.1 and a slope of 1.36.

Now the question to ask is, what happens if a point or two gets shifted around. Well if we go here and we shift this point around, move to there, look at that. It didn't have much impact. The slope is still 1.34. The r is 0.99, so it didn't do much. Move this one a little bit. Move the center one's a little bit and nothing much happens. With this one line started moving a little bit, but we're still at 0.98 for r.

Now what happens on the other hand, if we choose one of the extreme points. And this is when we're going to start worrying, and we start seeing things happen. Especially if we move it. Look at this. It's dragging the line with it. So if we move it radically—so for example, if we move it all the way down here then look at that. The correlation coefficient now is minus 0.27. So it's gone from a very high positive to negative.

And things get worse if I also move this one, so all of a sudden, we get a completely different picture of what's going on. So we're at r=-0.64. And you might say, oh well look you moved two of the five, so you'd expect this to happen. Well OK, what happens if we add some more points along the original line. And look at this, it's not having a tremendous effect.

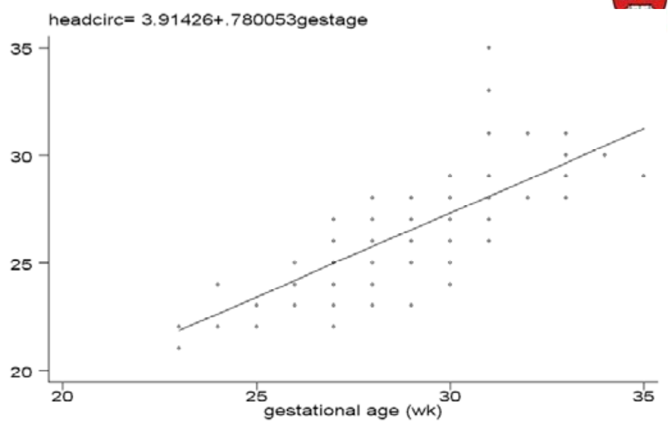
Yeah, it's having some effect, but look at that. So here we've got 17 points. So now two of the 17 are way off, but the other 15 of the 17 don't have enough influence to carry the day. So we can carry on like this if you want, and look at what it's going to take to shift the line around. It's going to take quite a bit to bring the line around. It still only a quarter of the way there. We're still at only 0.24.

Still just these two points, these two extreme points, carry a lot of influence. So that's the moral of this story. Now what I want you to do, is I would like for you to do this by yourself. The instructions are down here on the screen, and you can see how to do this. It is very easy. And there's the website down there too. All right.

Now, while you're here, we can also revisit that statement that I'd made to you, remember, if you had the v shaped, if you have y is equal to the absolute value of x. And I said, oh look, that's probably going to give you a correlation of 0. Well let's try that and see what happens when we fit a line here. Look at that, 0.05. So that bears out that statement that I made last week with a correlation coefficient.

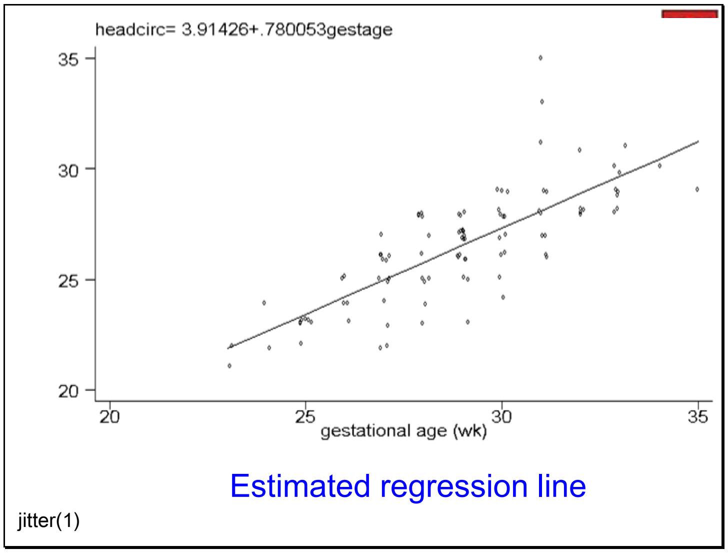
In fact, remember we looked at the fatal automobile accidents and what percentage had an alcohol content above 0.08. So we had age along the horizontal, and then we had the percentages on the vertical. And once again, there it is. The correlation coefficient is equal to 0.00. This is a little bit more extreme than what we had before, but you get the idea.

So go there and try this for yourself, and you'll get a little better feel for how the least squares line works. We could have stuck with what Boskovic did, but that's for the next course. The way you will learn about mean absolute deviation, or median regression, or L^1 regression it's called. There are more robust ways of doing regression, but go ahead and enjoy yourselves.



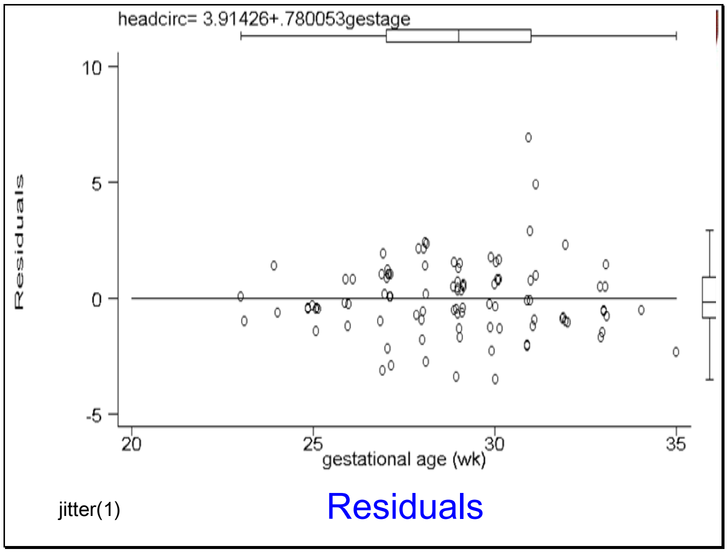
Estimated regression line

Returning to the original data—non-jittered—showing the estimated least squares regression line. In this case it looks like the straight line is not a bad fit. What if it were not? Before answering that question let us see what might lead us to such a conclusion.



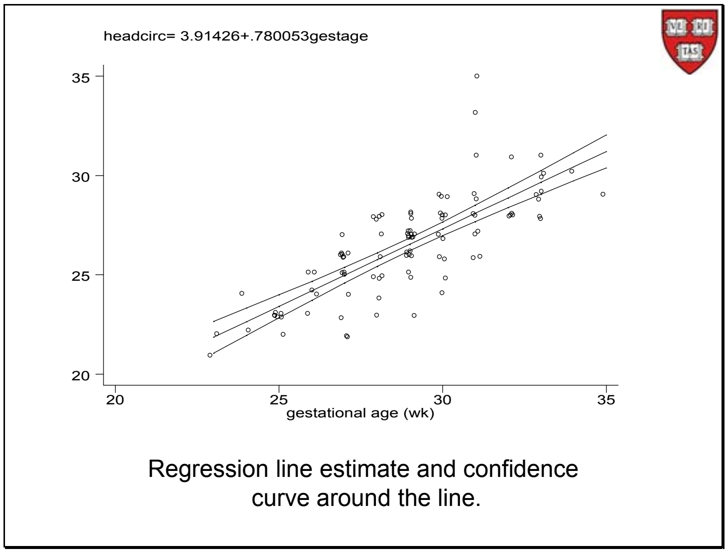
Estimated regression line

Let us look at a jittered version of the data, and that should not change our opinion that the relationship is linear.



Residuals

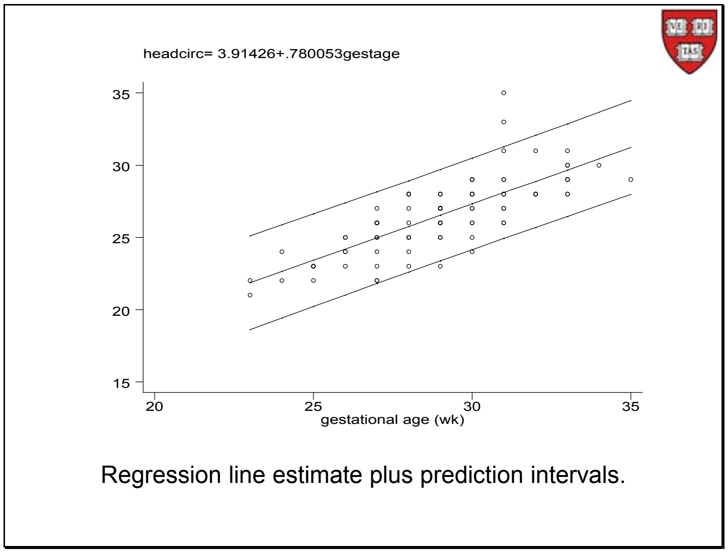
We can also subtract regression line from all the points thus creating the residuals. Now remember we said that we were assuming homoscedasticity for these. On looking at the graph we do not see any residual shape in these—they are not all negative at the left and positive at the right, for example—they look pretty much evenly distributed about zero, and their size is evenly distributed and thus our assumption of linearity and homoscedasticity seems well supported by the data.



Regression line estimate and confidence curve around the line.

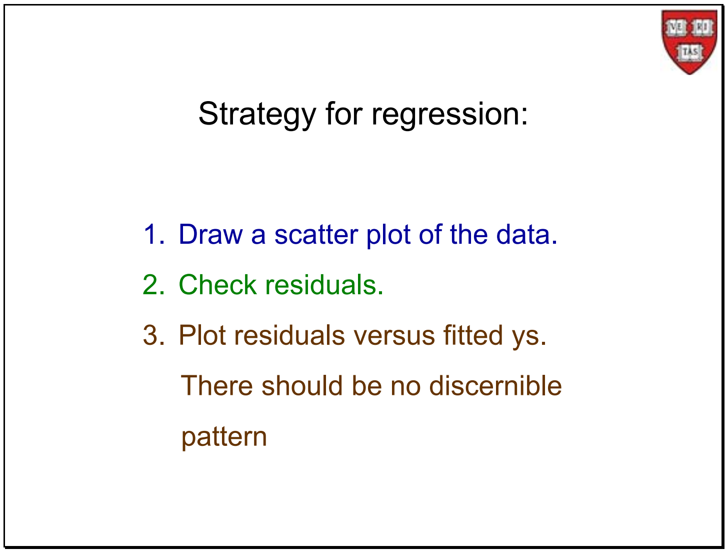
Remember the line is a statistic, it can vary from sample to sample. We can place a 95% confidence interval around this line that has the following interpretation: For any fixed gestational age, this interval serves as a 95% confidence interval for the mean (the regression line) of all kids with that gestational age at birth.

We see that the interval is tighter near the center than it is on the extremities, and that intuitively makes sense because most of our data are near the center, and so that is where we have most of our information.



Regression line estimate plus prediction intervals.

We might also be interested in where 95% of the babies will be, and that can be answered with this estimated prediction interval. For example, our least squares estimate of where 95% of the babies' head circumferences are for babies who are born at 25 weeks gestational age, then that is given by this graph when evaluated at 25 weeks.



Strategy for regression:

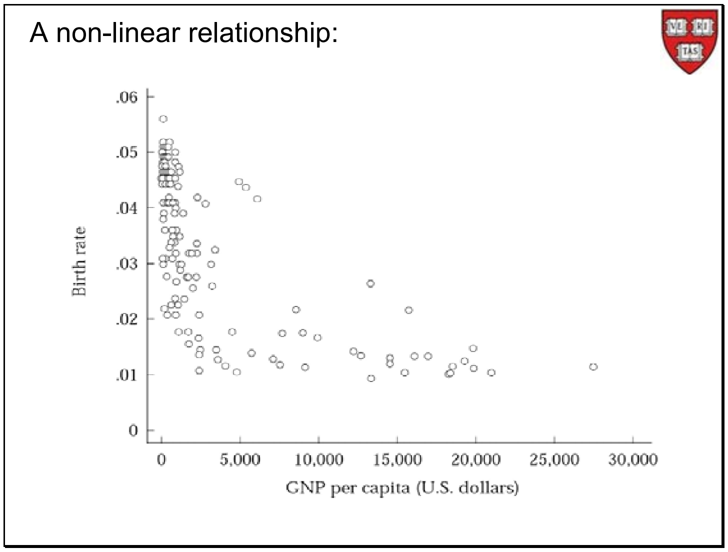
1. Draw a scatter plot of the data.
2. Check residuals.
3. Plot residuals versus fitted ys.

   There should be no discernible

   pattern

So in summary, what is our general strategy for simple linear regression? We start by drawing a scatter plot of the data. That will give us an idea of what the relationship is and whether we in fact have linearity or not. Then once we fit the model we should plot at our residuals to see if there are any discernible patterns. The patterns to look for are first to see if there are any trends going up or down or U-shaped etcetera that would reveal non-linearity. Second we should also look for patterns of non-homoscedasticity—for example are the residuals flute-shaped to denote more variability at one extreme or the other.

There should not be any discernible patterns. Also be on the lookout for outliers since they may overly affect your results, as you noticed when exercising with the applet above.



A non-linear relationship:

Here is an example of a scatter plot we would rather not see if we are about to fit a linear regression line. This plot shows the birth rate for some countries against their per capita gross national product (GNP) in US dollars. The relationship between the two is certainly not linear.

Some shapes we can rectify and transform the data to obtain a linear relationship. What do we mean by transform and what kind of transform to investigate. By transform we are talking about instead of looking at the measure directly, a transform of it might be more appropriate. For example, we all know about the inverse square law in physics, Newton's law of gravity is one example of this. It says that the intensity of some quantity is inversely proportional to the square of the distance from the source of that quantity. Now, if instead of distance, we invented a new quantity called tsidtsid, say, which is calculated by taking the reciprocal of the square of the distance, then the intensity would now be linearly related to tsidtsid. So we would have the tsidtsid law and we could draw a linear regression of intensity on tsidtsid.

Conversely, if instead of the intensity of the quantity we measure the reciprocal of the square root of the quantity, then it would have a linear relationship with the distance.

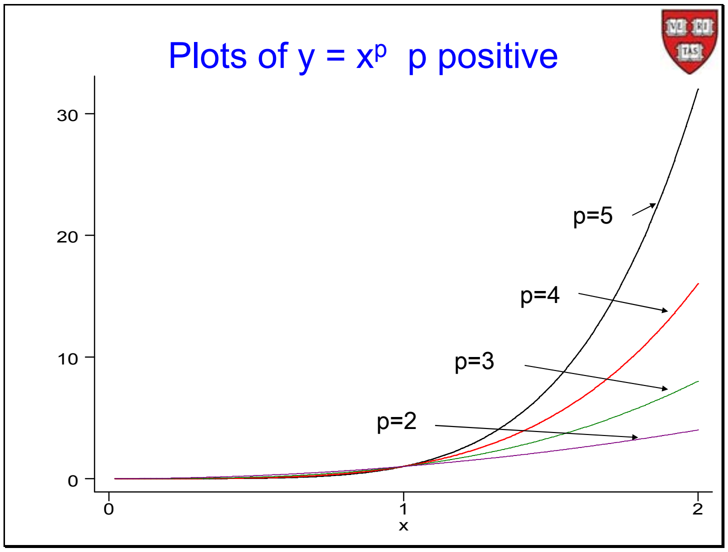
What all this is saying is if

$$I = \frac{1}{d^2} \quad \text{then if} \quad tsidtsid = \frac{1}{d^2} \quad \text{then} \quad I = tsidtsid.$$

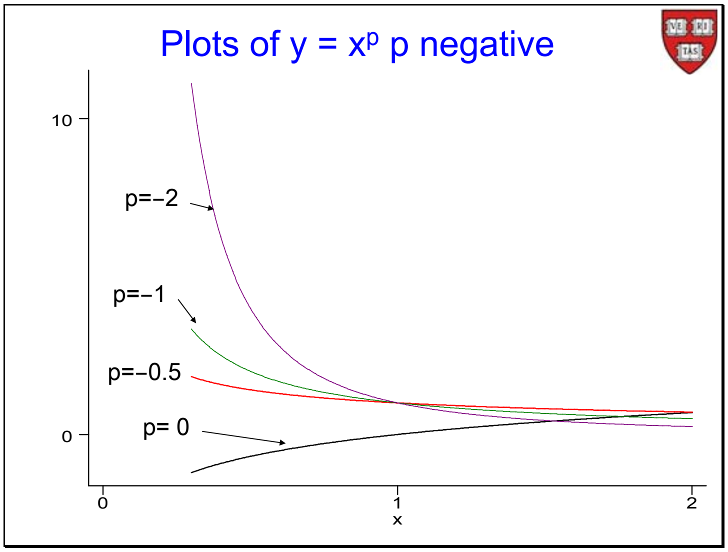
Or, alternatively, if above, then

$$\frac{1}{\sqrt{I}} = d.$$

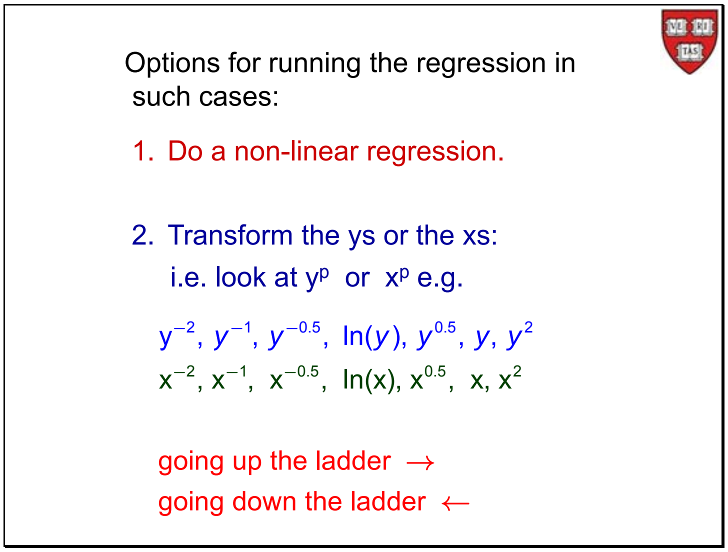
So sometimes by exploring scatter plots we can find a way to transform our data to achieve linearity.



Consider this family of curves. They show the relationship of $y=x^p$ for various positive values of p. As p gets larger, the rate at which y increases with x gets steeper and steeper. So if we take the inverse transformation, $x^{1/p}$, then that would slow the rate of increase of y, and in fact achieve linearity.



With the power p negative we get different shapes. Like this we can build up an armamentarium of transformations to bring to the problem of straightening out curves.



Options for running the regression in such cases:

1. Do a non-linear regression.

2. Transform the ys or the xs:
   i.e. look at y^p or x^p e.g.

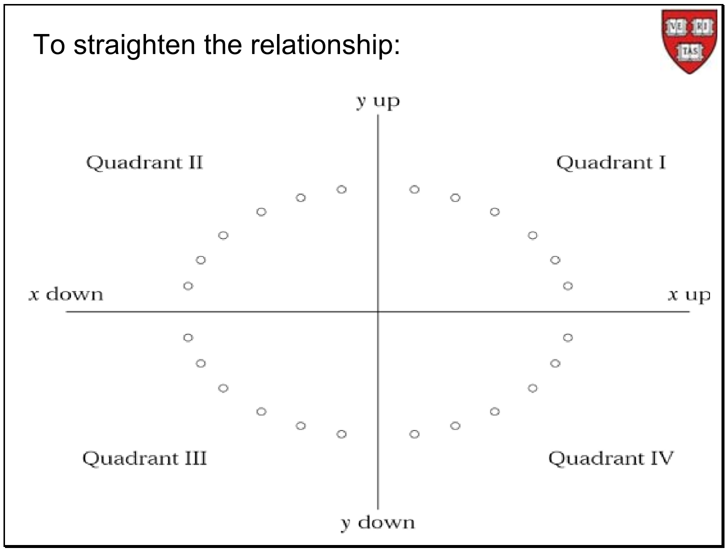
   y^{-2}, y^{-1}, $y^{-0.5}$, $\ln(y)$, $y^{0.5}$, y, y^2

   x^{-2}, x^{-1}, $x^{-0.5}$, $\ln(x)$, $x^{0.5}$, x, x^2

going up the ladder →

going down the ladder ←

We can create a ladder of transformations starting on the left with a large negative value for p and moving right by increasing the value of p. To fill out the ladder we can replace p=0 with the logarithm. That way we can search up and down the ladder to see if anything straightens out our relationship to allow us to fit a linear regression.



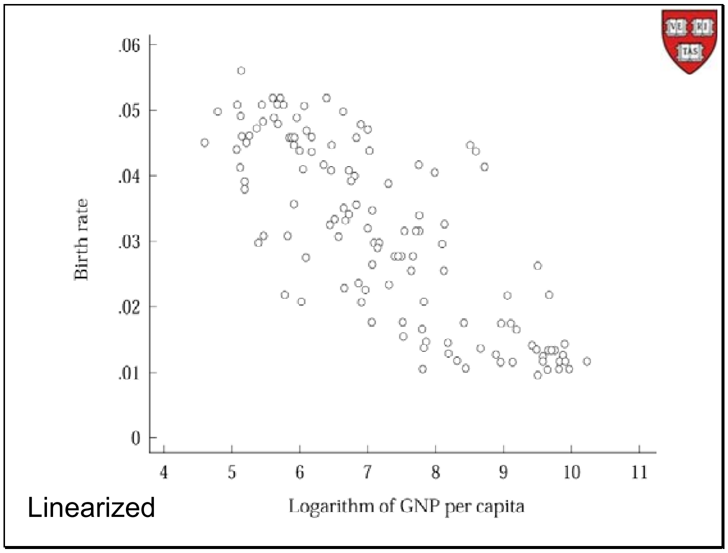
To straighten the relationship:

To guid us in our search of the appropriate transformation we have this quadrant guide due to John Tukey[12] The way we use it is we start with the power of y is zero and the power of x is zero. Then identify the shape of the scatter plot by seeing which quadrant it falls into. If the

[12] Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. Also look at http://onlinestatbook.com/2/transformations/tukey.html

shape that best describes the scatter plot is that in Quadrant I, then either up the power on x, and or up the power on y to achieve linearity. Quadrant II would suggest upping the power on y and or lowering the power on x. And so on.

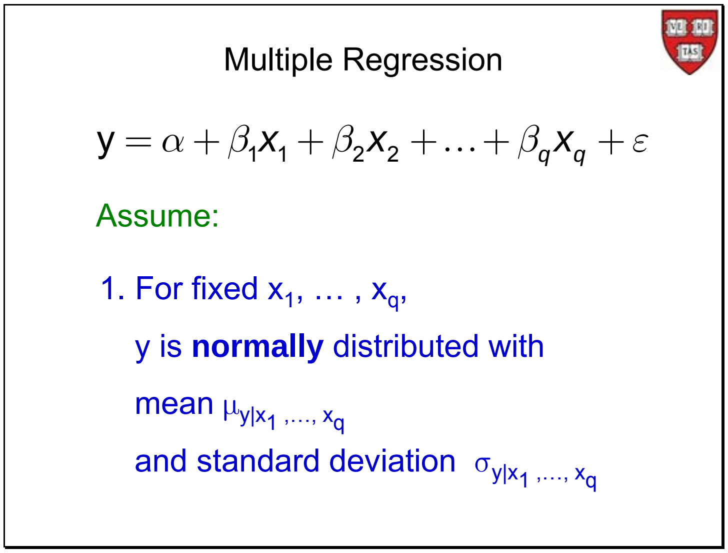
Why upping the power on y might be different than lowering the power on x might be because of the error structure that best describes the situation at hand.



Returning to the birth rate – GNP relationship above, we might identify that with the shape in Quadrant III, and that suggests looking at a decrease in the power of either x or y. In this case, replacing y with its logarithm, results in this scatter plot which reveals a relationship which is much closer to linearity.

Sometimes these transformations work and sometimes they do not, but they are always worth a try.

Multiple Linear Regression



Multiple Regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \varepsilon$$

Assume:

1. For fixed $x_1, \ldots , x_q$,

   y is **normally** distributed with

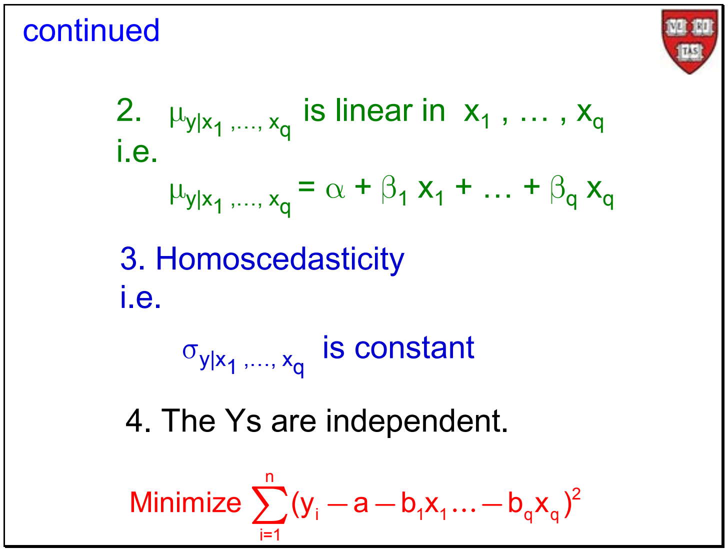
   mean $\mu_{y|x_1 ,\ldots, x_q}$

   and standard deviation $\sigma_{y|x_1 ,\ldots, x_q}$

All too often a single explanatory variable is not sufficient to fully explain the changes in the distribution of a response variable in which case we might introduce more explanatory variables. To handle these extra variables, let us consider our assumptions just like we did before.

We still assume normality of our response variable, except that now we want that this normality to be true for a fixed collection of xs. So instead of just gestational age we might also want to include the babies' birth weights, the mothers' ages and so on, bring in a number of factors.

Now fix all those, and we shall have a number of ys, just like before, and we are going to assume that they are normally distributed, with a particular mean, and those means are going to lie on a plane. Around those means we are going to have standard deviations.



continued

2. $\mu_{y|x_1,\ldots,x_q}$ is linear in $x_1, \ldots, x_q$
i.e.

$$\mu_{y|x_1,\ldots,x_q} = \alpha + \beta_1 x_1 + \ldots + \beta_q x_q$$

3. Homoscedasticity
i.e.

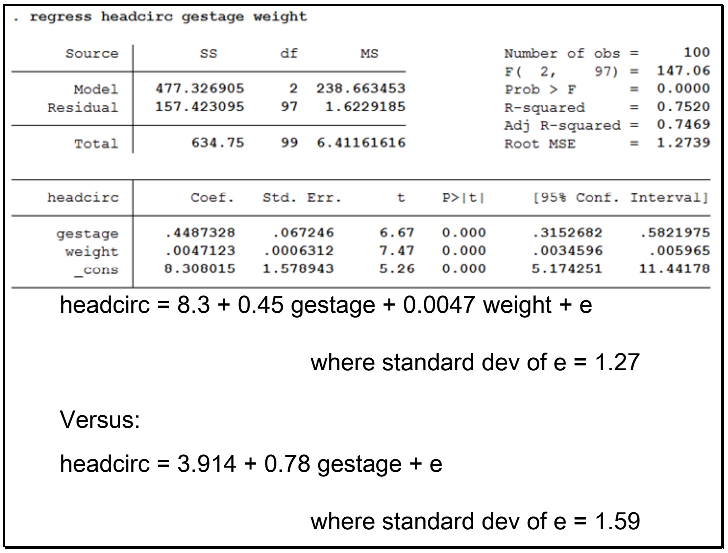
$\sigma_{y|x_1,\ldots,x_q}$ is constant

4. The Ys are independent.

Minimize $\sum_{i=1}^{n}(y_i - a - b_1x_1 \ldots - b_qx_q)^2$

We assume that the mean is the linear in all xs, and whereas before we fit a straight line, now we are going to fit a plane. And we are going to retain our assumption of homoscedasticity. Once again, the ys are independent. So it's the same assumptions as before, except we're talking about a plane instead of a line. Not a big change.

Once again, we, or rather, Stata, are going to minimize the sum of squares of residuals.



```
. regress headcirc gestage weight

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =  147.06
       Model |  477.326905     2  238.663453           Prob > F      =  0.0000
    Residual |  157.423095    97   1.6229185           R-squared     =  0.7520
-------------+------------------------------           Adj R-squared =  0.7469
       Total |      634.75    99  6.41161616           Root MSE      =  1.2739

------------------------------------------------------------------------------
    headcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gestage |   .4487328    .067246     6.67   0.000     .3152682    .5821975
      weight |   .0047123   .0006312     7.47   0.000     .0034596     .005965
       _cons |   8.308015   1.578943     5.26   0.000     5.174251    11.44178
------------------------------------------------------------------------------
```

headcirc = 8.3 + 0.45 gestage + 0.0047 weight + e

where standard dev of e = 1.27

Versus:

headcirc = 3.914 + 0.78 gestage + e

where standard dev of e = 1.59

Here is an example of multiple regression. The Stata command is as before except we now add another explanatory variable to the list, namely *weight*; the birth weight of the babies. Let us see if adding this extra explanatory variable helps us or not in understanding head circumference.

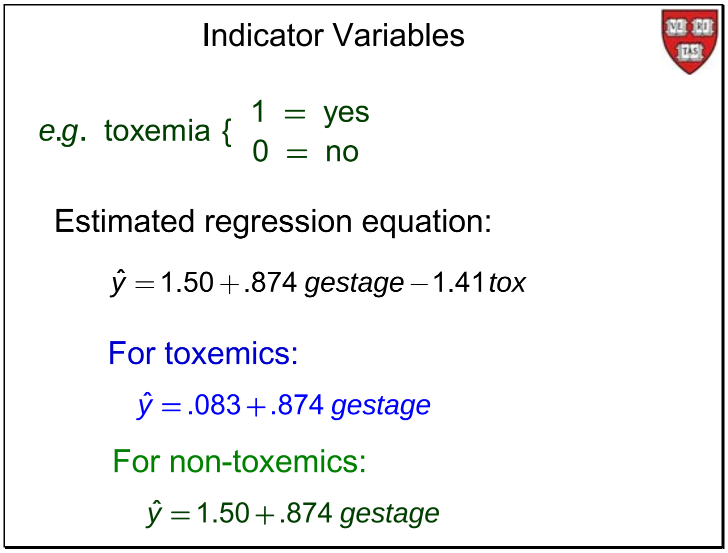
We see the two regression lines, the first without birth weight and the second with birth weight added. The coefficient for gestational age has decreased some in the presence of birth weight. The interpretation of this is that in the first equation gestational age, which is correlated with birth weight, was carrying the explanation afforded by both, whereas with birth weight in the equation it can do its own work. Once again, by looking at both, the p-value and confidence interval associated with birth weight we see that it is a significant explanatory variable. We also see that the standard deviation of the residuals has gone down from 1.59 to 1.27 centimeters, a 20% decrease. So it seems that when looking at head circumference of a baby, both its gestational age and its weight at birth are important.

Note that the coefficients are not pure numbers, they both depend on the uniots of measurement being utilized. For example, the 0.45 coefficient of gestational age has units of cms/week, and the 0.0047 coefficient for weight is for cms/gram. Had the babies weights been measured in Kilograms, then the coefficient would have been 4.7 cms/Kg.

The other statistic to keep an eye on is the R-squared. It went from 0.6 to 0.75, a good improvement.

The main problem with multiple regression is that the way we have done the modeling here is we have acted as if these two explanatory variables, gestational age and weight, are additive. Is it possible that there is some interaction between gestational age and weight? Once we have opened the Pandora's box the two variables, and later with even more, may also interact with each other in a complicated way. We return to that but first introduce another class of variables.

Indicator Variables



Indicator Variables

e.g. toxemia $\left\{ \begin{array}{l} 1 = \text{yes} \\ 0 = \text{no} \end{array} \right.$

Estimated regression equation:

$$\hat{y} = 1.50 + .874\, gestage - 1.41\, tox$$

For toxemics:

$$\hat{y} = .083 + .874\, gestage$$

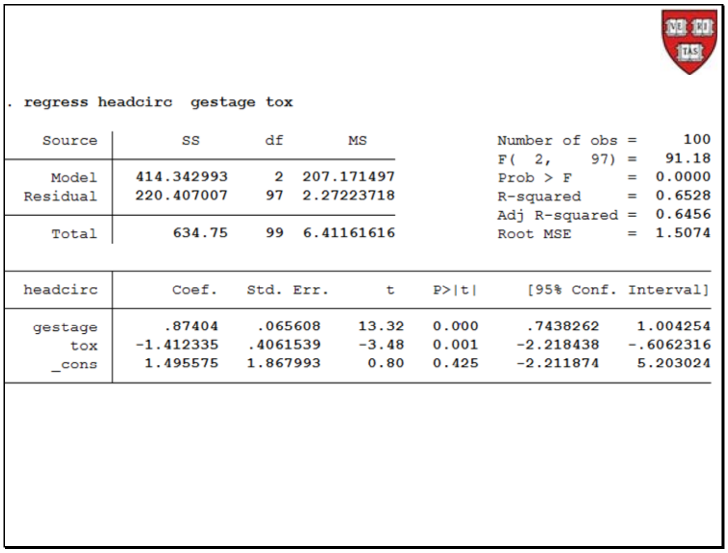
For non-toxemics:

$$\hat{y} = 1.50 + .874\, gestage$$

An important set of variables we can introduce as explanatory variables are the dichotomous variables. They are the ones we have been using to build tables. In the regression context they are called indicator variables. The ideas we are about to develop extend in a straightforward manner to categorical variables also, but let us just look at the 2-valued, simple ones. Sometimes in the literature you will see these called dummy variables, but why inflict that on them.

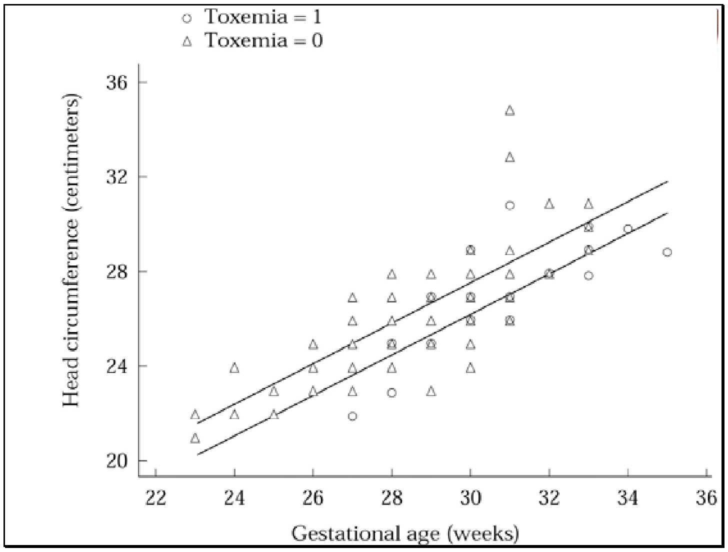
For example, let us look at toxemia. If the mother was toxemic at the time of delivery, then we let the *tox* variable take the value 1. If the mother was not, then this variable will take on the value 0. Above we see the result of fitting a regression line with the variables gestational age and toxemia.

We can view this as two regression lines, one when the mother was toxemic and one when she was not. Above we see the definitions of these two lines, and we note that they are parallel with different intercepts.



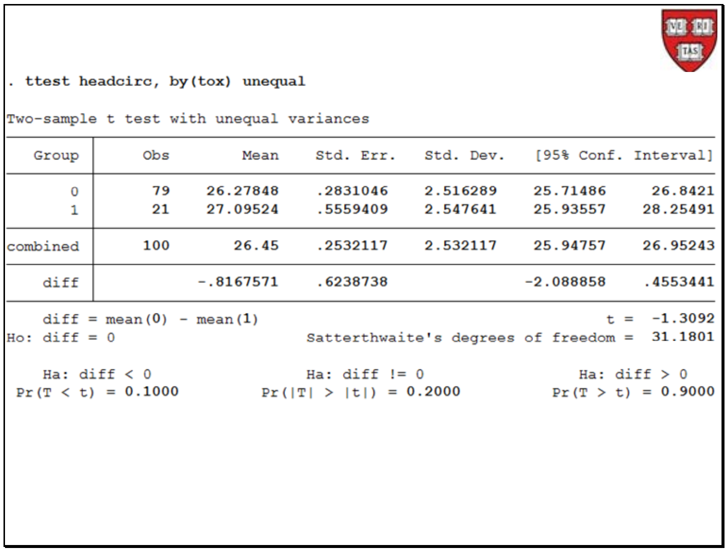
```
. regress headcirc  gestage tox

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =   91.18
       Model |  414.342993     2  207.171497           Prob > F      =  0.0000
    Residual |  220.407007    97  2.27223718           R-squared     =  0.6528
-------------+------------------------------           Adj R-squared =  0.6456
       Total |      634.75    99  6.41161616           Root MSE      =  1.5074

------------------------------------------------------------------------------
    headcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gestage |     .87404    .065608    13.32   0.000     .7438262    1.004254
         tox |  -1.412335   .4061539    -3.48   0.001    -2.218438   -.6062316
       _cons |   1.495575   1.867993     0.80   0.425    -2.211874    5.203024
------------------------------------------------------------------------------
```

We obtained all this information from this Stata output. The indicator variable denoting toxemia is significant, although the R-squared is not as good as it was when we introduced birthweight into the regression.



Here is a plot of the two parallel lines. What these say is that an extra week of gestational age as far as the head circumference goes has the same average effect whether the mother suffered from toxemia or not. They start from different starting points, but the effect of an extra week is the same.

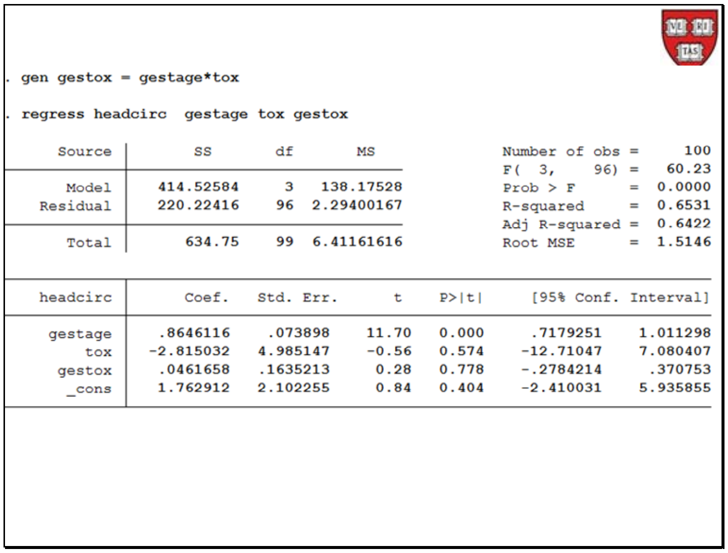
We can investigate whether this makes physiological sense or not.



```
. ttest headcirc, by(tox) unequal

Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      79    26.27848    .2831046    2.516289    25.71486     26.8421
       1 |      21    27.09524    .5559409    2.547641    25.93557    28.25491
---------+--------------------------------------------------------------------
combined |     100       26.45    .2532117    2.532117    25.94757    26.95243
---------+--------------------------------------------------------------------
    diff |           -.8167571    .6238738               -2.088858    .4553441
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -1.3092
Ho: diff = 0                     Satterthwaite's degrees of freedom =  31.1801

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.1000         Pr(|T| > |t|) = 0.2000          Pr(T > t) = 0.9000
```

We can perform a t-test between the two groups of infants and find we cannot reject the null hypothesis of no difference in head circumference even though we did find a difference when we included birth weight in the model.

Toxemia happens later in the pregnancy. We saw that later in the pregnancy, as measured by gestational age, does have an impact on head circumference. Also, birth weight increases with gestational age. So how all of these explanatory variables work together is quite complex. Is there an interaction between these explanatory variables?

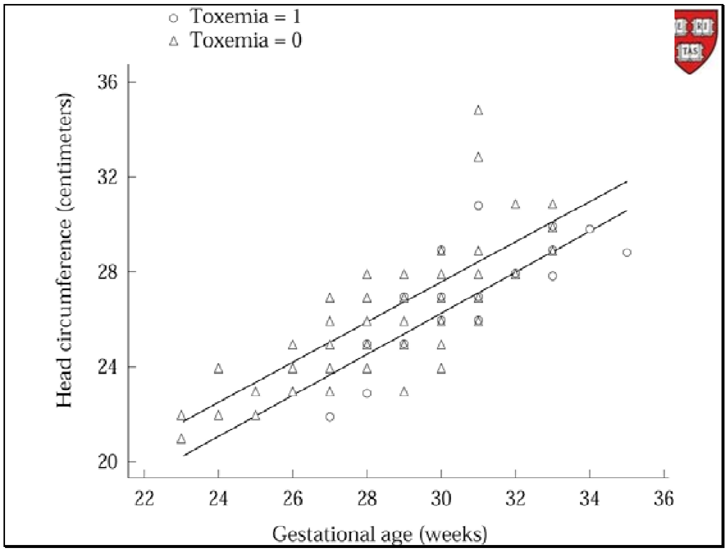


```
. gen gestox = gestage*tox

. regress headcirc  gestage tox gestox

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =   60.23
       Model |  414.52584     3  138.17528           Prob > F      =  0.0000
    Residual |  220.22416    96  2.29400167           R-squared     =  0.6531
-------------+------------------------------           Adj R-squared =  0.6422
       Total |     634.75    99  6.41161616           Root MSE      =  1.5146

------------------------------------------------------------------------------
    headcirc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     gestage |   .8646116    .073898    11.70   0.000     .7179251    1.011298
         tox |  -2.815032   4.985147    -0.56   0.574    -12.71047    7.080407
      gestox |   .0461658   .1635213     0.28   0.778    -.2784214     .370753
       _cons |   1.762912   2.102255     0.84   0.404    -2.410031    5.935855
------------------------------------------------------------------------------
```

We can generate a new variable called gestox which is equal to gestational age times toxemia. This is called an interaction term. It should tell us whether the babies born of mothers who suffered from toxemia act differently from those whose mothers did not have toxemia. In the latter case gestox=0, whereas it is not zero in the former. That means that the two regression lines will no longer be parallel.

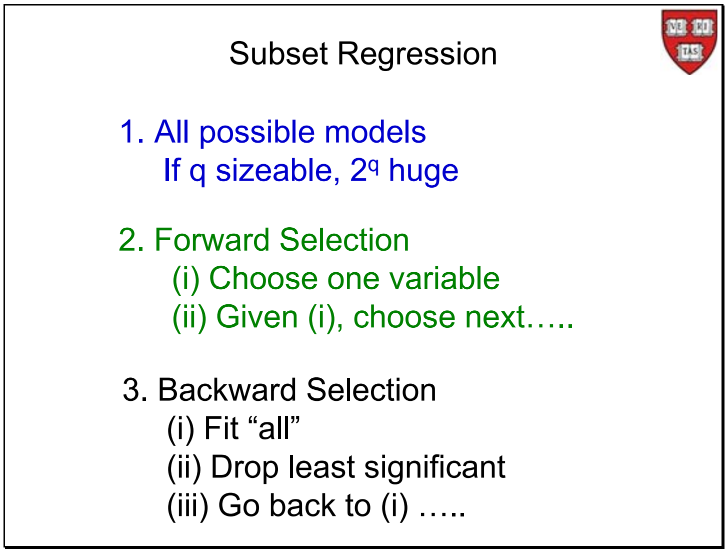
Once we let this variable enter into the model we see that now toxemia is no longer important. It is not important either by itself or as part of the interaction term. The only explanatory variable left of any import is gestational age.

The R-squared has gone down to 0.64. This should not be happening, we should be improving the explanatory capability of a model by adding explanatory variables, not decreasing the capability.



If we plot the regression lines now, they are almost, but not quite, parallel. We have run into a problem of collinearity we explore below. This is what can happen when we explore which variables to include in our regression model when we have the choice of many variables, as we do in this small data set, even. Not only do the variables represent different explanatory aspects of the outcome variable in question, they may also interact with each other in a way that makes the model fitting much more complex.

Subset Regression



Subset Regression

1. All possible models
   If q sizeable, 2^q huge

2. Forward Selection
   (i) Choose one variable
   (ii) Given (i), choose next…..

3. Backward Selection
   (i) Fit “all”
   (ii) Drop least significant
   (iii) Go back to (i) …..

Which explanatory variables and in what form to include into our regression model is called the subset regression problem. All too often in a study the problem being studied is serious and patients, and their families, are being asked to give of themselves to be studied mostly to help future patients so we do not wish to miss something important. Add on top of that the scientists' favorite theories and as a result there are usually a large number of candidates to be entertained as possible explanatory variables. On the other hand, if this study is to be of use to future patients it has to be generalized and we do not want to have results that are peculiar only to this particular set of patients. That usually argues for a parsimonious model with a few, choice, and important explanatory variables.

There are three main strategies for attacking the subset regression problem. The general problem has not yet been solved, so we proceed by relying very much on our own personal experience with model fitting.

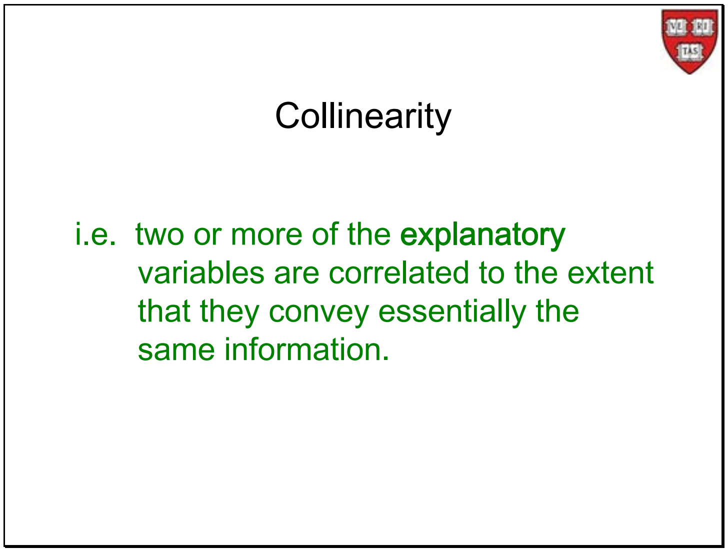
One approach is to investigate all possible models. If there are q possible explanatory variables that means there are 2^q possible models, ignoring interaction terms, that can be fit. If the number of observations is large, then each one of these models may take some time to fit, and if on top of that, q is sizable then you could spend the rest of your life, like Tycho Brahe, fitting models to data. This approach also creates a huge Bonferroni problem, if you will, of related tests.

We could rank the models by looking at the respective R-squareds, and that is one way to proceed. Not advisable, but a way.

Another approach is what is called the forward selection approach. Here what you do is find the best single variable somehow. The one with the highest R-squared, or the highest t-value, and then force that into the model. Then with that first on in the model look for the next explanatory variable to include, and then the third, and so on. The problem with this approach is that once you have the first two, there is no guarantee that that is the best two-explanatory variable model.

Finally there is the backward selection approach. In this one you fit all the explanatory variables into the model and then work your way down by eliominating the least impoirtant ones, one by one. Once again, there is no guarantee that when you end up with a model with m explanatory variables, it is the best m explanatory variable model.

So here are three approaches that you can take, and there are some computer programs that have been automated to do all this. I would strongly recommend that you do not use any of these. They are problematic. For one, it is quite probable thatyou do not get the same answer from all three approaches, and it is often very difficult to understand why.



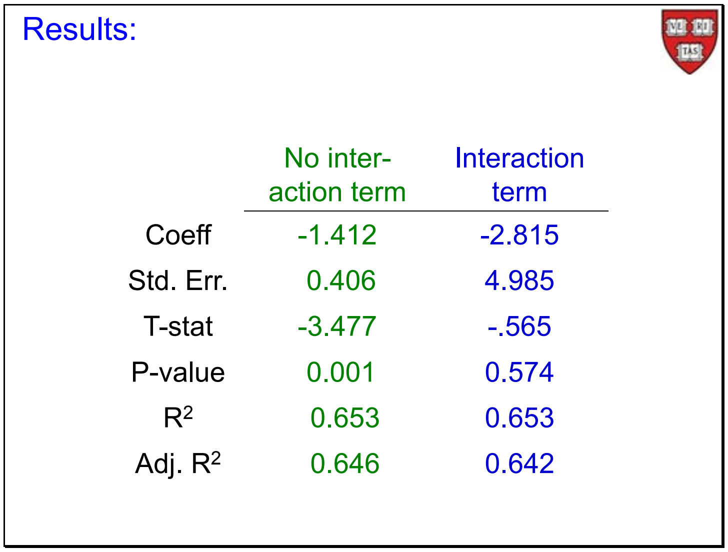
Collinearity

i.e. two or more of the **explanatory** variables are correlated to the extent that they convey essentially the same information.

One of the problems is that you run into is what is called collinearity. The most extreme case of collinearity occurs if you had somebody's weight measured in kilograms in one variable, and in another variable that person’s weight measured in pounds. If you plot those two variables against each other the relationship would be a straight line—thus collinear. When two variables are collinear then you really only have one variable, not two, the second variable does not add any more information that is not contained in the first variable. The two variables are perfectly correlated with each other.

It can get even more complicated than that. You might have three variables who are collinear in the sense that two of them can predict the third one exactly. And so on, with multiple variables.

And it gets even more complex when the variables are not perfectly correlated with each other, but almost perfectly correlated, or highly correlated with each other. You might still then run into interpretation and fitting problems. And that is the collinear problem.



Results:

	No inter-action term	Interaction term
Coeff	-1.412	-2.815
Std. Err.	0.406	4.985
T-stat	-3.477	-.565
P-value	0.001	0.574
R^2	0.653	0.653
Adj. R^2	0.646	0.642

That is the problem we have run into when looking at toxemia. If you look at the two models, one with no interaction term and then the one with the interaction term, and you looked at the toxemia coefficient. In the first model it is -1.442. In the second model it almost doubled in value to -2.815. Plus, when you look at the standard error associated with this term, it increased tenfold. In the meantime the p-value went from 0.001 to 0.574 and there was no change in the R^2. Such a radical change in these statistics, brought about by the introduction of a single explanatory variable (accompanied by no change in the R^2) is an indication of instability; exactly the sort of instability brought about by collinearity.

In this case, these unstable results make sense, because we know that toxemia tends to happen later in the pregnancy, and thus toxemia is related with higher gestational age. So the introduction of the interaction term between toxemia and gestational age is attempting to introduce three very interrelated terms in to the regression. Take care when this happens.