Objectives

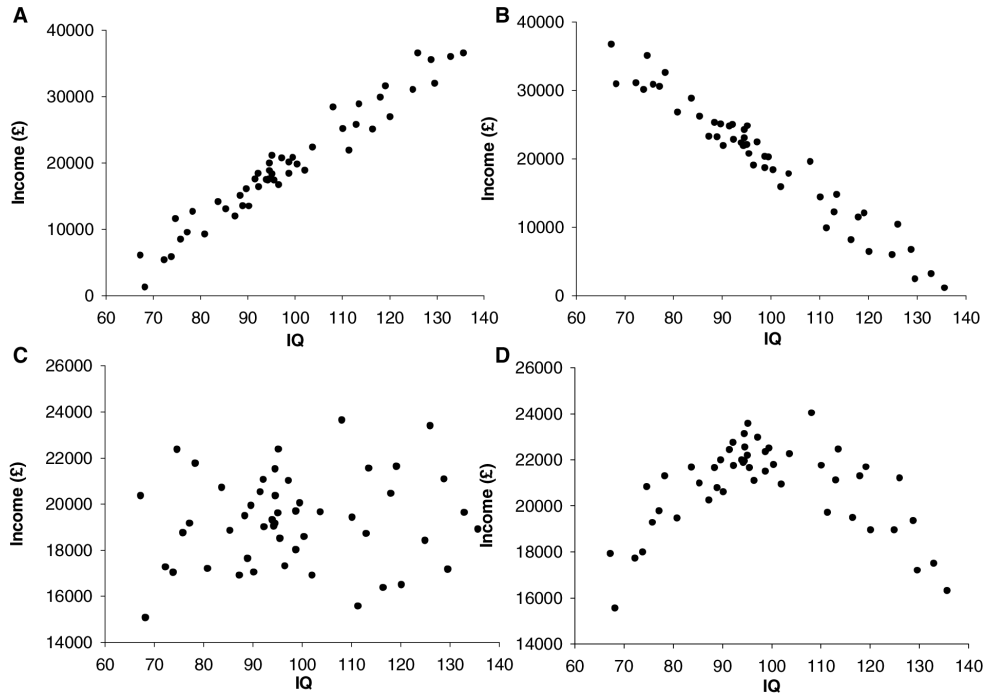
We'll examine two ways to examine the relationship between two variables — correlation and regression. They're conceptually very similar.

Stuff with a solid edge, like this, is important.

But remember — you can totally ignore stuff with single / double wavy borders.

2.1 Scatter plots

Suppose you measure two things about a group of subjects — IQ and income, say. How can we establish if there's any relationship between the two? The first thing to do is to draw a **scatter plot** of the two variables. To do this, we take one of our variables (e.g. IQ) as the *x* axis, and the other as the *y* axis. Each subject is then plotted as one point, representing an {IQ, income} pair. This might show us any of several things:



Fictional scatterplots. ***A: positive correlation*** *between IQ and income. As IQ goes up, income goes up.* ***B: negative correlation*** *between IQ and income. As IQ goes up, income goes down.* ***C: no correlation*** *between the two.* ***D:*** *there's a relationship, but it's not a straight line (it's not a* ***linear*** *relationship). People with high IQs and people with low IQs both earn less than those with middling IQs.*

It's always worth plotting the data like this first. However, for our next trick we'd like a statistical way to work out **if** there's a relationship, **how big** it is, and **in what direction** it goes. Please note that we'll only talk about ways to establish things about a **linear relationship** between two variables; if it's non-linear (e.g. the bottom right figure), it's beyond the scope of this course.

2.2 Correlation

We will call the degree to which they are related the **correlation** between the two variables. If *Y* gets bigger when *X* gets bigger, there's a **positive correlation;** if *Y* gets smaller when *X* gets bigger, there's a **negative correlation;** if there's no linear relationship, there's a **zero correlation.** Here's how we work it out.

The covariance

First, we need some sort of number that tells us how much our two variables vary together. Let's suppose we have *n* observations. Let's call our two variables *X* and *Y*. We first find the two means, $\bar{x}$ and $\bar{y}$. Then we can calculate something called the **sample covariance:**

$$\mathrm{cov}_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}$$

Look at the first part of the equation first — it's very like the sample variance (if we changed all the *y*s to *x*s in this equation, we'd have s_X; if we changed all the *x*s to *y*s, we'd have s_Y). (And yes, if you're wondering, if we wanted the population covariance, we'd divide by *n* rather than *n*–1, but we don't.)

Perhaps you can see from the equation how it works. For a given {*x*, *y*} point, if *x* is very far above the *x*-mean ($\bar{x}$), and *y* is very far above the *y*-mean ($\bar{y}$), then a big number gets added to our covariance. Similarly, if *x* is very far below the *x*-mean ($\bar{x}$), and *y* is very far below the *y*-mean ($\bar{y}$), then a big number gets added to our covariance. Both these occurrences suggest a positive linear relationship (like the top-left part of our figure). On the other hand, if *x* is very far above $\bar{x}$, and *y* is very far below $\bar{y}$, then a large *negative* number gets added to our covariance; the same's true if *x* is very far below $\bar{x}$ and *y* is very far above $\bar{y}$. Points near the mean don't tell us so much about the relationship between *x* and *y*, and they don't contribute much to the covariance score. If there's no relationship between *X* and *Y*, then when *x* is above $\bar{x}$, about half the time *y* will be above $\bar{y}$ and the covariance will get bigger, but about half the time *y* will be below $\bar{y}$ and the covariance will get smaller, so the covariance ends up being about zero.

The Pearson product–moment correlation coefficient, r

The covariance tells us how much the two variables are related, but it has a problem — the actual value of the covariance depends on the standard deviations of our two variables as well as the correlation between them. A covariance of 140 might be an high correlation if the standard deviations are small, but a poor correlation if the standard deviations are large. We can get round this problem by calculating *r*:

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{s_X s_Y}$$

It turns out that *r* varies from –1 (perfect negative correlation), through 0 (no correlation), to +1 (perfect positive correlation).

Incidentally, the correlations in our picture were +0.79 (figure A), –0.81 (figure B), 0.10 (figure C), –0.04 (figure D).

'Zero correlation' doesn't imply 'no relationship'

That should be immediately apparent: I've just told you that the correlation between IQ and income in figure D was –0.04, nearly zero, and yet there's clearly a very strong relationship — it just isn't a linear one. **Always plot your data** to avoid drawing mistaken conclusions from *r* values.

Correlation does not imply causation

Finding that X and Y are related does not mean that X and Y are **causally** related. It's easy to jump to this assumption if the relationship is plausible — we might intuitively think that clever people get better jobs, for example, and thus accept a positive correlation between IQ as income as indicating causation. It doesn't. Maybe the causal relationship is backwards: having more money might improve your IQ. Maybe the two are connected through a third variable: Z causes X and Z causes Y (e.g. maybe having rich parents means you're more likely to have a high IQ, and

also makes you more likely to get a well-paid job as an adult). The point is, we just can't tell from the plain correlation.

Adjusted r

If we measured IQ and income for a *sample* of just two people — let's say {IQ 110, £20,000} and {IQ 120, £25,000} — and calculate *r*, we'll find that there's a perfect correlation, +1. If you plot only two points on a scatterplot, you can always join them perfectly with a straight line. This doesn't mean that the correlation is +1 in the *population!* So there's something slightly wrong with our sample correlation statistic, *r* — it's a *biased estimator* of the **population correlation,** which we write as ρ (Greek letter rho). We can to do something to make it a better **(unbiased)** estimator. We can calculate the **adjusted *r*, r_{adj}**:

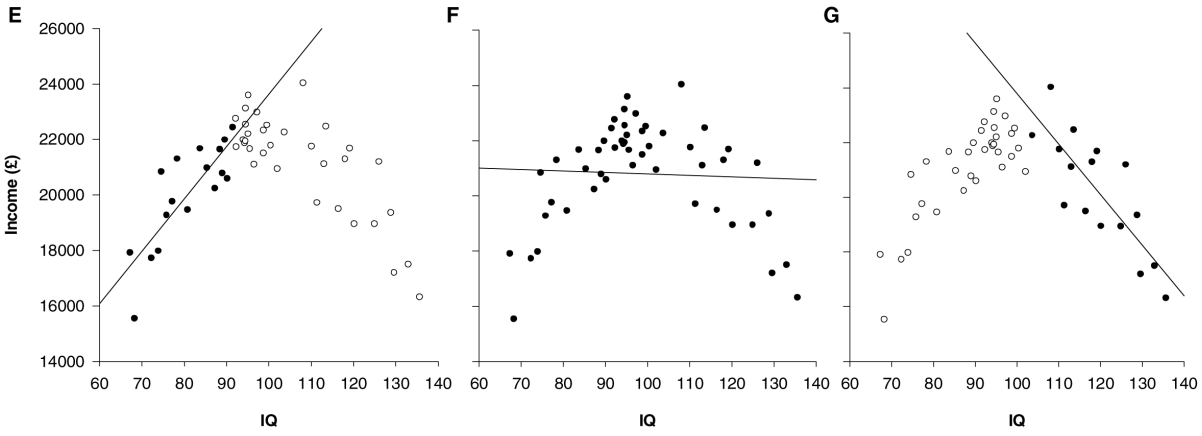
$$r_{adj} = \sqrt{1 - \frac{(1-r^2)(n-1)}{n-2}}$$

If the sample size is large, *r* and r_{adj} will be about the same. Please note that doing this will give you a *positive value* for r_{adj}, since square roots can't be negative… so you need to look at the original data or *r* value to work out *which way* (+ or –) the correlation should be.

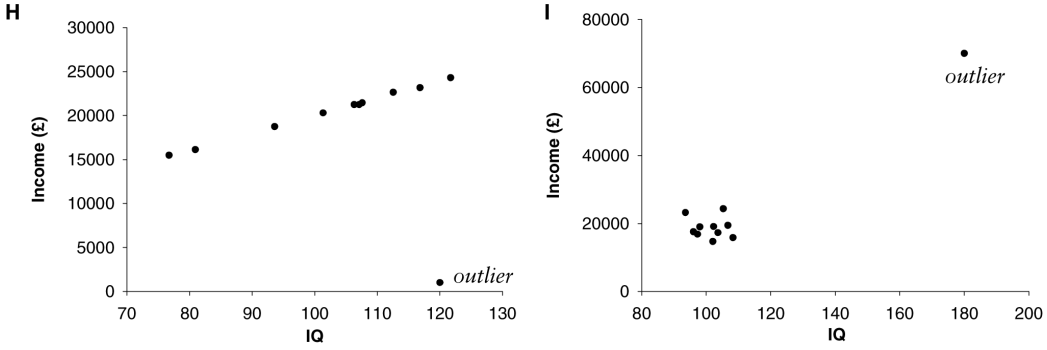
Beware if your correlation is based on a restricted range of data

It's obvious that if we sample too few data points, we won't get a very good estimate of *r* for our population (that's what calculating r_{adj} is meant to sort out). But it should also be clear that if we sample from a **restricted range,** we can also get the wrong answer, even if we sample many observations within that restricted range. Here's an extreme example (figures E–G below): depending on the range of data we sample, we can contrive to find a negative, zero, or positive correlation between our two variables.

Beware outliers



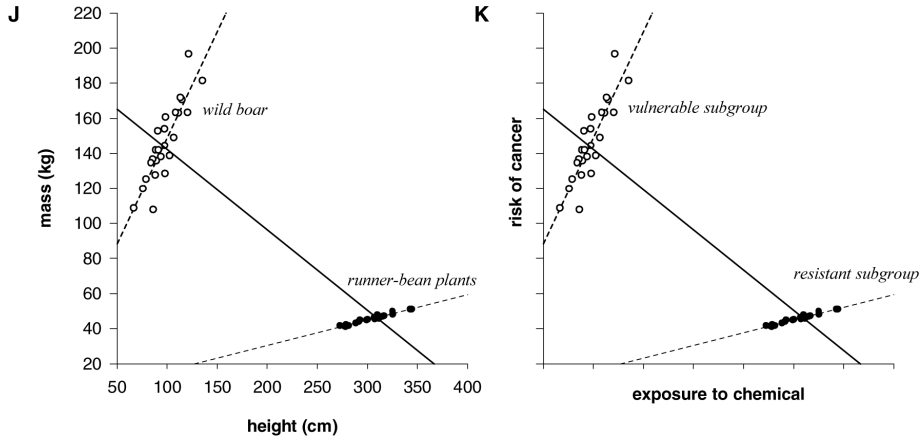
Above: *Sampling a restricted range of data can overestimate r* ***(E)*** *or underestimate r* ***(G)*** *compared to sampling the whole range* ***(F).*** *Black dots are part of the sample; white dots are part of the population that wasn't sampled. The straight line represents the correlation.* ***Below:*** *Outliers can have large effects on r. In* ***(H)*** *the outlier makes r nearly 0; without it, r would be nearly 1. In* ***(I),*** *the outlier makes r nearly 1; without it, r would be nearly 0.*



Extreme values, or *outliers*, can have large effects on the correlation coefficient. (We won't talk about what to do with them in the IB course, but you should be aware of the problems they can cause.) Two examples are shown in figures H–I.

Beware if your population has distinct subgroups

We can also encounter problems if our measurements aren't from one homogeneous population. A couple of examples are shown in figures J–K (but subgroup effects can be a good deal more subtle than this!).



*Fictional data illustrating problems with subgroups. **(J)** Correlation between height and weight for various things we found in a magic forest. If we measure an overall correlation, we may find that tall things weigh less (negative correlation between height and mass), but this is only because we have two very different subgroups. But we have **heterogeneous subsamples** — within each subgroup (wild boar and runner beans) there is a positive correlation. **(K)** A less stupid example. If we are investigating whether something is carcinogenic, we might find a negative correlation, suggesting (if we have designed our experiment so that we know that the chemical **caused** any observed change in cancer rate) that the chemical protects from cancer. But we must check, because this could be due to a subgroup effect: a more detailed analysis may reveal a vulnerable subgroup (who get high rates of cancer) and a resistant subgroup (who aren't as likely to get cancer); in this example, the rates of cancer are actually increased by the chemical in both subgroups.*

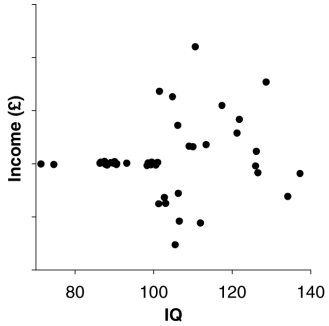
2.3 Is a correlation 'significant'?

Assumptions we must make

If all we want to do is to *describe the sample* that we have (e.g. with correlation and / or regression), we don't have to make any assumptions — although correlation and regression both aim to describe a **linear relationship** between two variables, so if the relationship isn't linear, then the answers we get from correlation and regression won't *mean* very much.

But if we want to perform statistical tests with the data (e.g. 'is this correlation coefficient significantly different from zero?'), we will effectively be asking questions to do with the *underlying population* that our sample was drawn from (i.e. 'what is the chance that a sample with correlation r came from an underlying population with correlation $\rho = 0$?'). This requires making some assumptions, or our statistical tests won't be meaningful. Basically, the data shouldn't look too weird:

- The variance of Y should be roughly the same for all values of X. This is often called *homogeneity of variance;* its opposite, what you don't want, is called *heteroscedasticity* (Greek *homo* same, *hetero* other, *skedastos* able to be scattered).



Heteroscedasticity: a Bad Thing. The variance in income is very different for low-IQ and high-IQ data.

- If we are asking questions about ρ, we must assume that both X and Y are normally distributed.
- For all values of X, the corresponding values of Y should be normally distributed (and vice versa). [You may see the last two assumptions referred to together as the assumption of 'bivariate normality'.]

Testing the 'significance' of r — is r significantly different from zero?

Let's suppose we take a sample of people, measure their IQs and incomes, and correlate them to find r. That's the correlation in the sample; but is there a correlation in the whole population? Our null hypothesis is that the population correlation coefficient (ρ) is zero. Without going into the details, we can compute a number called a ***t* statistic:**

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

We can use this number, t, to perform a ***t* test** with n–2 degrees of freedom. (The statistical catchphrase is that the number we have just calculated is 'distributed as t with n–2 degrees of freedom'; the t distribution is much like the normal distribution that we've mentioned before, so we need to look up the probability corresponding to our t statistic just like we might look up a probability corresponding to a Z score.) To interpret this using tables, we can look up the **critical value of *t*** for our particular value of α and the number of degrees of freedom; if our t statistic is bigger than the critical value, it's 'significant' and we reject the null hypothesis that there's no correlation in our underlying population.

(Note that we use r, not r_{adj}, for this test.)

This is an example of a t test; we'll cover these properly in Practical 2.

2.4 Spearman's correlation coefficient for ranked data (r_s)

If our X and Y data are both **ranked** (see below for how to rank data), we can calculate the correlation coefficient r just as normal, except that we'll call it $\boldsymbol{r_s}$ (sometimes called Spearman's rho). However, when we want to test the significance of r_s, we have a problem, because we cannot make our assumption that the data are normally distributed. Some argue that there are substantial problems inherent in computing the significance of r_s (see Howell, 1997, p. 290). Anyway, with these caveats, what we'll do is to look up **critical values of $\boldsymbol{r_s}$** if $n \leq 30$, and if $n > 30$ we'll calculate t and test that, just as before:

$$t_{n-2} = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \text{ where } n > 30$$

To answer the question **'are these values in a particular order?'** you can correlate the rank of the data with the rank of their position. For example, suppose you take large spoonfuls of bran flakes from the top of a cereal packet, one by one, and find the mean weight of individual bran flakes in each spoonful. These weights, in milligrams and in order, are 70, 84, 45, 50, 48, 40, 38, 40, 25, 30. If you want to establish whether it's true that big bran flakes come out of the packet first, you can correlate the set of positional ranks {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} with the corresponding ranks of the data {9, 10, 6, 8, 7, 4.5, 3, 4.5, 1, 2} to get $r_s = -0.918$ ($p < .001$).

How to rank data

Suppose we have ten measurements (e.g. test scores) and want to rank them. First, place them in ascending numerical order:

5	8	9	12	12	15	16	16	16	17

Then start assigning them ranks. When you come to a tie, give each value the mean of the ranks they're tied for — for example, the 12s are tied for ranks 4 and 5, so they get the rank 4.5; the 16s are tied for ranks 7, 8, and 9, so they get the rank 8:

X:	5	8	9	12	12	15	16	16	16	17
rank:	1	2	3	4.5	4.5	6	8	8	8	10

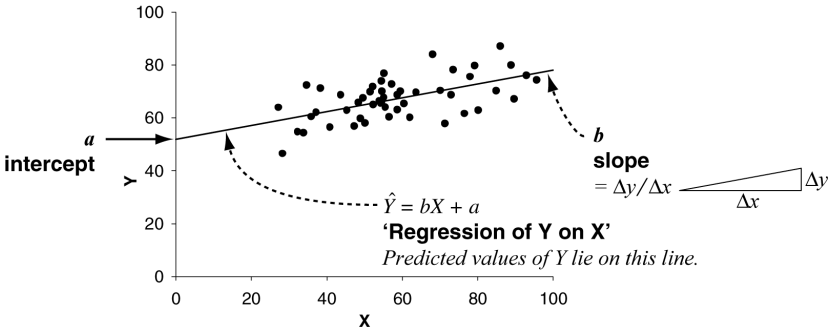
2.5 Regression

We've used correlation to measure *how much* of a relationship there is between two variables. We can use a related technique, **regression,** to establish exactly *what* that relationship is — specifically, to make predictions about one variable using the other. Suppose there's a positive correlation between serum cholesterol in 50-year-old men and their chance of having a heart attack in the next five years. If Mr Blobby has a serum cholesterol twice that of Mr Slim, are his chances of having a heart attack doubled? Increased by a factor of 1.5? Tripled? Let's find out.

If we call our two variables *X* (cholesterol) and *Y* (chance of having a heart attack), we can write an **regression equation** that describes the **linear relationship** between *X* and *Y*. It's just the equation of a straight line:

$$\hat{Y} = bX + a$$

We call this the **regression of Y on X,** meaning that we're predicting *Y* from *X*, not the reverse. The *Y* with a 'hat' ($\hat{Y}$) just means 'the predicted value of *Y*'. This is the picture that this equation represents:



The regression equation and what it means. You might also see it written $y = y_0 + ax$, *or some other equivalent.*

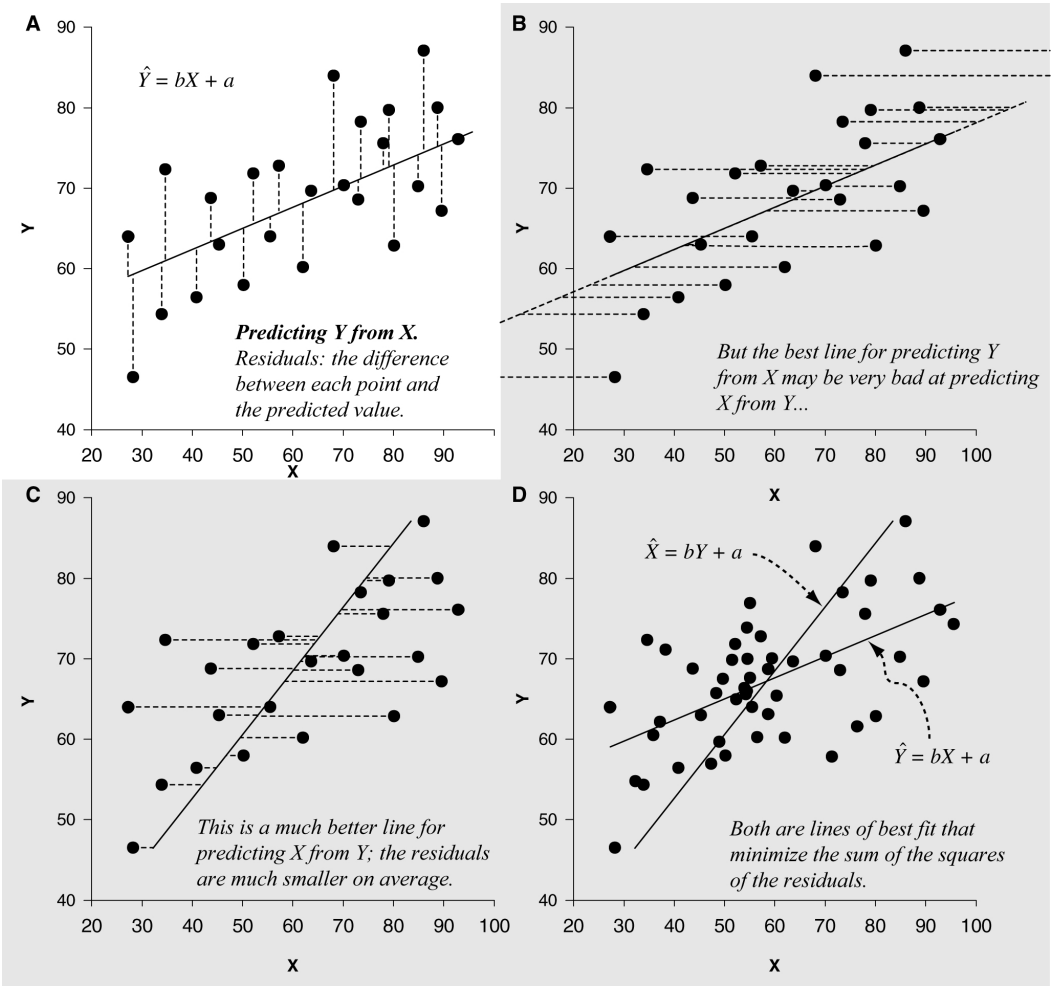
We could draw thousands of lines like this. So which one is the **best fit** to our data? If we take a particular line $\hat{Y} = bX + a$, then for each {*x*, *y*} point, we can calulate a predicted value $\hat{y} = bx + a$. From this, we can calculate how wrong our prediction was: the prediction error is $y - \hat{y}$. This error is often called the **residual,** because it's what you have left after you've made your prediction. Since this will sometimes be positive and sometimes be negative, we can square it to get rid of the +/– sign, giving us the *squared error:* $(y - \hat{y})^2$. So we should aim to find a line that gives us the minimum possible total prediction error, or *sum squared error,* $\sum (y - \hat{y})^2$. (This procedure is called **least squares regression.**) As it happens, this is when

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\text{cov}_{XY}}{s_X^2} \quad \text{(or } b = r\frac{s_Y}{s_X} \text{ if that's easier with your calculator)}$$

Note that regression is not a symmetrical process: the best-fit line for predicting *Y* from *X* is probably not the same as the best-fit line for predicting *X* from *Y* (illustrated in the figure below). This is **different from correlation,** which doesn't 'care' which way round *X* and *Y* are.

To save you the bother of doing this by hand, **your calculator should give you A and B (regression) and *r* (correlation) directly.** Learn how to use it for the exam! Typically, you put it into 'statistics mode' or 'linear regression' mode, clear the stats



Residuals and lines of best fit. ***(A)*** *What's a residual?* ***(B–D), which are LESS IMPORTANT,*** *show why predicting Y from X is different from predicting X from Y.*

memory, then enter each data point as an {*x*, *y*} pair — then you can read out the answers.

Plotting and interpreting the regression line

To plot the line, you just need any two {*x*, $\hat{y}$ } pairs — though it helps if they're far apart, because this makes your line more accurate, and it's often wise to plot a third point somewhere in the middle to make sure it lies on the same line! The line will also pass through the points {0, *a*} and { $\bar{x}$, $\bar{y}$ }.

If you actually need to predict a *y* value from some *x* value — say your father's got a particular cholesterol level and you wanted to predict his risk of a heart attack — then you can just use the regression equation directly. **Beware of extrapolating beyond the original data, though.** If you've based your regression equation on 50-year-old men with a cholesterol level of 4–8 mM, they may be pretty useless at predicting heart attack risks in 50-year-old men with a cholesterol level of 12 mM, or 100-year-old men, or 50-year-old women. Within the range of your data, though, you can also make statements like 'for every 1 mM drop in cholesterol, one would expect a 10% reduction in the risk of a heart attack' (or whatever it is); this information is based on the **slope** of the regression line.

Finally, remember that **correlation and regression do not necessarily represent causation** (see above).

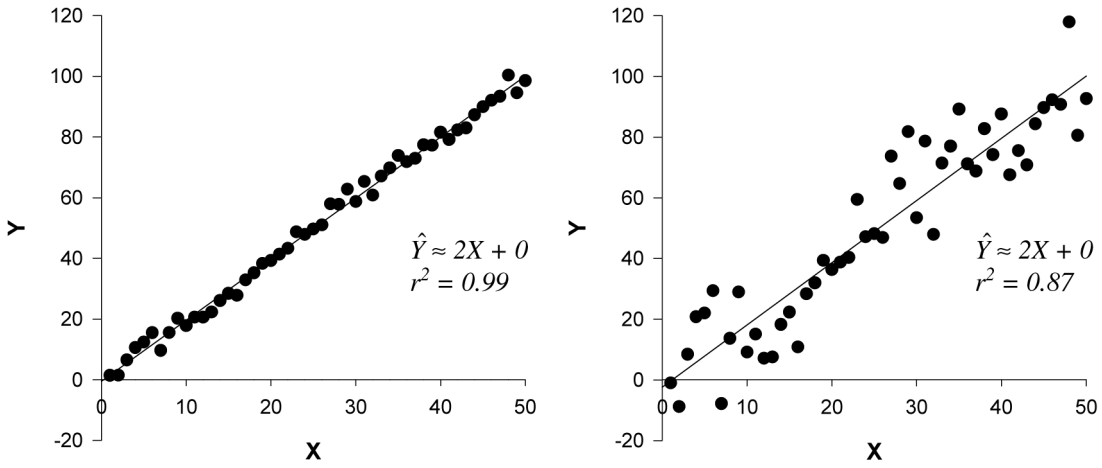
r^2 as a measure of how good a correlation or regression is

So far, we've drawn a regression line. But how good it is at predicting *Y* from *X* depends on how much of a relationship there is between *Y* and *X* — we could draw a

regression line where *Y* was the chance of having a heart attack and *X* was shoe size, but it wouldn't be a very good one. How can we quantify 'how good' our best fit is?

r^2 represents the proportion of the variability in *Y* that's predictable from the variability in *X*, or (equivalently) the proportion by which the error in your prediction would be reduced if you used *X* as a predictor. Let's say the correlation between cholesterol levels and heart attack risk were ridiculously high, at $r = 0.8$; then $0.8^2 = 0.64 = 64\%$ of the variability in the risk of heart attacks would be attributable to variations in cholesterol. If $r = 0.1$, then $0.1^2 = 0.01 = 1\%$ of the variability in the risks of heart attacks would be attributable to differences in cholesterol levels.

Note, once again, that this doesn't tell you anything about causality. If rainfall is predictable from twinges in your gammy knee, that doesn't necessarily mean that twinges cause rain, or that rain causes twinges.



Two regressions with nearly identical equations ($\hat{Y} = 2X$) but different values of r^2.

Mathematical statement of this property of r^2

Let's start by taking the worst-case scenario. If you knew *nothing* about your subject's cholesterol level (*X*), how accurately could you predict his risk of a heart attack (*Y*)? Your best guess would be the mean risk of a heart attack, $\bar{y}$, and your error would be described in some way by the standard deviation of *Y*, s_Y, or the variance s_Y^2. The variance, remember, is

$$s_Y^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$$

Now the bottom part of that, $n - 1$, is the number of **degrees of freedom (*df*)** our estimate of the variance was based on. (This was in the first handout — if you have *n* numbers, and you use them to calculate the sample mean, $\bar{y}$, then you can subsequently only alter $n - 1$ of the numbers freely without altering the mean. This is called the number of *degrees of freedom* you have left — it is the number of *independent* observations on which a given estimate is based.) The top part is the sum of the squares of the deviations of *Y* from the mean of *Y*, which we shorten to the **sum of squares of *Y* (SS_Y).** So we can write the variance as

$$s_Y^2 = \frac{SS_Y}{df}$$

Let's now suppose that we do know our subject's cholesterol. We have a whole set of *n* observations with which to calculate a regression line, i.e. *a* and *b*. (Since we calculate two numbers, we're left with $n - 2$ degrees of freedom in our data.) But now we can estimate our subject's heart attack risk rather better, we hope — and the error in doing so will be related to the *residuals* (error) of our regression's prediction. This thing is called the **residual variance,** or **error variance:**

$$s_{residual}^2 = \frac{\sum (y - \hat{y})^2}{n - 2} = \frac{SS_{residual}}{df}$$

and its square root, $s_{residual}$ (sometimes written $s_{Y \cdot X}$ to show that *Y* has been predicted from *X*), is called the **standard error** of the estimate (it's like a standard deviation — the square root of the variance of the errors is the standard deviation of the errors, abbreviated to the **standard error**). Well, rather than work all this out by hand, let's use somebody else's result:

$$s_{residual}^2 = s_Y \sqrt{(1-r^2)\frac{n-1}{n-2}} = s_Y \sqrt{1-r_{adj}^2}$$

Actually, it's generally easiest to do the calculations in terms of the sums of squares, not variances, because then we don't have to worry about all these degree-of-freedom corrections (*r* and r_{adj} and this $n - 1$, $n - 2$ business) — you can't add two variances together unless they're based on the same number of degrees of freedom, but **you can add sums of squares together** any way you like — and we find that

$$SS_{residual} = SS_Y(1-r^2)$$

$$r^2 = \frac{SS_Y - SS_{residual}}{SS_Y}$$

$$SS_Y = SS_Y(r^2) + SS_{residual}$$

In other words, the total variability in *Y* is made up of a component that's related to *X* ($SS_Y \cdot r^2 = SS_Y - SS_{residual}$, which we can also write as $SS_{\hat{Y}}$, the variability in the *predicted* value of *Y*) and a component that's residual error ($SS_{residual}$). Translated to our cholesterol example, people vary in their cholesterol levels (SS_X), they vary in their heart attack risk (SS_Y), a certain amount of the variability in their heart attack risk is predictable from their cholesterol ($SS_{\hat{Y}}$), and a certain amount of variability is left over after you've made that prediction ($SS_{residual}$). Or,

$$SS_Y = S_{\hat{Y}} + SS_{residual}$$

where

$$r^2 = \frac{SS_{\hat{Y}}}{SS_Y}$$

2.6 Advanced real-world topics

As with all the wavy-line sections, this section certainly isn't intended to be learned! It's for use with real-world problems that you may encounter. You will not be tested on any of this in the exam.

What's 'regression to the mean'?

Something related to regression, but quite interesting. It was discovered by Galton in 1886. He measured the heights of lots of families, and calculated the 'mid-parent height' (the average of the mother's and the father's height) — call it *X* — and the heights of their adult children — call it *Y*. He found that the average mid-parent height was $\bar{x}$ = 68.2 inches; so was the average height of the children ($\bar{y}$ = 68.2 inches). Now, consider those parents with a mid-parent height of 70–71 inches; the mean height of their children was 69.5 inches — the height of these children (69.5) was closer to the mean of *all* the children ($\bar{y}$ = 68.2) than the height of the parents (70–71) was to the mean of all the parents ($\bar{x}$ = 68.2). But this wasn't a genetic phenomenon, it was a statistical phenomenon, and it worked backwards: if you took children with a height of 70–71 inches, the mean mid-parent height of their parents was 69.0 inches. This is called **regression to the mean.**

Why does it happen? Suppose we have the variables *X* and *Y*, with standard deviations s_X and s_Y, and the correlation between them is *r*. We've previously seen that

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y}$$

and the regression slope *b* is

$$b = \frac{\mathrm{cov}_{XY}}{s_X^2}$$

Therefore,

$$slope = b = r\frac{s_Y}{s_X}$$

So a change of one standard deviation in *X* is associated with a change of *r* standard deviations in *Y*. And we know the regression line always goes through the point at the means of both *X* and *Y* — that is, the point $\{ \bar{x}, \bar{y} \}$. Therefore, unless there is perfect correlation ($r = 1$), the predicted value of *Y* is always fewer standard deviations from its mean than *X* is from its mean. Remember that predicting *Y* from *X* is different from predicting *X* from *Y*, unless the two are perfectly correlated? This is another way of saying the same thing.

Examples of regression to the mean (from Bland & Altman, 1994, BMJ 309: 780)

- If we are trying to treat high blood pressure, we might measure blood pressure at time 1, then treated, and then measured again at time 2. We might see that blood pressure goes down *most* in those who had the *highest* blood pressure at time 1, and we might interpret this as an effect of the treatment. We'd be wrong; this is regression to the mean. It would happen even if the treatment had no effect. The two sets of observations (time 1, time 2) will never be perfectly correlated (because of measurement error and biological variation); $r < 1$. So if the difference between our 'high blood pressure' subgroup and the whole population was *q* at time 1, it will be *rq* at time 2 — i.e. the difference from the population mean will have shrunk. We should have compared our treated group to a randomized control group.
- In one study, people reported their own weight and had their weight measured objectively. A regression was used to predict reported weight from measured weight; the regression slope was less than 1. This might lead to you interpret that very fat people underestimate their weight when they report it, and very thin people overestimate it. But we'd never have expected *perfect* correlation. All this might be is regression to the mean — and if we'd predicted measured weight from reported weight, we'd also have a slope less than one, from which we might have concluded the opposite: that very fat people overestimated their weights and very thin people underestimated them.
- When scientific papers are submitted to journals, referees criticize them and editors select the 'best' ones to publish on the basis of the referees' reports. Because referees' judgements always contain some error, they cannot be perfectly correlated with any measure of the true quality of the paper. Therefore, because of regression to the mean, the average quality of the papers that the editor accepts will be less than he thinks, and the average quality of those rejected will be higher than he thinks.

Partial correlation — dealing with the effects of a third variable

Sometimes we are interested in the relationship between two variables and know that a third variable is also influencing the situation. Imagine we examine the correlation between IQ (*X*) and income (*Y*), and find it to be positive, but we suspect that one reason that higher IQ predicts higher income is because people with higher IQs are more likely to get into university, stay for higher degrees, and so on — and it's the degree that gets you the higher income, not your IQ itself. So is there any *further* relationship between IQ and income once you've taken into account this effect of studying for longer? One way of investigating this is to look at the correlation between IQ (*X*) and (say) number of years of study (*Z*), and the correlation between income (*Y*) and number of years of study (*Z*). We can then calculate the **partial correlation** between IQ (*X*) and income (*Y*) *having taken account of the relationship of each of these to number of years of study.* We call this 'partialling out' the effects of number of years of study. We term the partial correlation coefficient between *X* and *Y* with the effects of *Z* partialled out $r_{xy.z}$, and calculate it like this:

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

Let's use some fictional numbers to illustrate this: suppose that the correlation between IQ and income is $r_{xy} = 0.6$, the correlation between IQ and years of study is $r_{xz} = 0.8$, and the correlation between income and years of study is $r_{yz} = 0.7$. Then the correlation between IQ and income having partialled out the effect of years of study would be only $r_{xy.z} = 0.09$. This would mean that $r_{xy.z}^2 = 0.0081$, so only 0.8% of the variability in income is predictable from IQ once you've taken account of the number of years of study, even though $r_{xy}^2 = 0.36 = 36\%$ of the variability in income is predictable from IQ. This would suggest that nearly all the ability of IQ to predict income was due to the fact that high IQs predict more years of study.

The point-biserial correlation (r_{pb}) for a dichotomous variable

If we ask the question 'is body weight correlated with sex?', we have a bit of a problem with assuming that 'sex' is normally distributed; it clearly isn't. Body weight probably is, but mammals are either male or female; the sex variable is **dichotomous** (Greek *dikhotomos*, from *dikho-*, in two; *temnein*, to cut).

No problem: simply assign two values to the dichotomous variable as you see fit — e.g. male = 0, female = 1 (or male = 56, female = 98; it doesn't matter at all). Then calculate *r* as normal. Officially this *r* is called r_{pb}, the point-biserial correlation coefficient, but you can treat it like any *r*, and test it for significance in the same way (a *t* test on $n - 2$ *df*) as we saw before:

$$t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

You might think that asking 'does weight vary with sex' and calculating a correlation is a bit daft here, and the more natural question is 'do the sexes differ in body weight?' You'd be right, really, but it is actually the same question. Since it's the same question, there must be a simple relationship between r_{pb} and the *t* statistic:

$$r_{pb}^2 = \frac{t^2}{t^2 + df}$$

The use of this is that if you test the difference between two groups (e.g. male body weights and female body weights) using a *t* test (which we'll cover in Practical 2), you can calculate r^2, and therefore the proportion of the variability in body weight explained by sex. And if you read the results of a *t* test in a research article, you can interpret them in terms of r^2 using this technique.

Correlations when the dichotomous variable is 'artificial'…

The male / female dichotomy is natural; all subjects are either one or the other. Sometimes a dichotomy is arbitrary, such as 'pass / fail' in an exam with a 60% pass mark; this dichotomy classifies people who scored 59% and people who scored 1% in the same category, but classifies people who scored 59% and people who scored 60% in different categories. If you have data like these and want to calculate a correlation, you have to use a slightly different technique; this is described by Howell (1997, p. 286).

Correlations with two dichotomous variables

If you want to calculate the correlation between two variables when *both* are dichotomous, again, you can do it. All you do is calculate *r* in the normal way; this time, its special name is ϕ (phi). And it's exactly equivalent to doing a χ^2 test (which we'll cover in Practical 4). And there's a relationship between the two:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Why is this useful? Again, because r^2 is a measure of the proportion in the variability in one variable that's explained by a variable — the practical significance of the relationship — and therefore so is ϕ^2. So if you see a χ^2 test reported in an article, you can calculate ϕ^2 to see whether the relationship is *important* (large) as well as *significant*; see Howell (1997, p. 285).

Is a regression slope (b) significantly different from zero?

We saw earlier how to test if a correlation (r) was significantly different from 0. Since correlation and regression are much the same thing, we can also calculate the t statistic from the regression parameter b (from $\hat{y} = bx + a$) using a different fornula. For this it helps to use the notation $s_{Y \cdot X}$ rather than $s_{residual}$ for the standard deviation of the residuals left over when we have predicted Y from X. We would find that the t statistic we've just worked out could also be found like this:

$$t_{n-2} = \frac{b \cdot s_X \sqrt{n-1}}{s_{Y \cdot X}}$$

As before, this t statistic is distributed with $n - 2$ degrees of freedom (which is why the subscript on the t is $n - 2$).

If we calculate two regressions, are they significantly different?

Suppose we calculate the relationship between smoking and life expectancy in males and females. We'd probably find that the more you smoke, the shorter you live ($b < 0$). Let's suppose we find that this relationship is stronger in males (e.g. $b_{\text{male}} < b_{\text{female}} < 0$), suggesting that males decrease their life expectancy more than females for a given increment in the amount they smoke (though, of course, the regression by itself doesn't tell you anything about causality). Is this difference between males and females significant? If we have two variables X_1 and X_2 that both predict a third variable Y, and two sample regression coefficients b_1 and b_2, then we can calculate a t statistic (with $n - 4$ *df*) for the null hypothesis that the two underlying population regression coefficients are the same:

$$t_{n-4} = \frac{b_1 - b_2}{\sqrt{\frac{s^2_{Y \cdot X_1}}{s^2_{X_1}(n_1 - 1)} + \frac{s^2_{Y \cdot X_2}}{s^2_{X_2}(n_2 - 1)}}}$$

'I have two values of r from different (independent) groups. Are they different?'

If we want to do the same with correlations rather than regressions ('is the correlation r_1 between male smoking and male life expectancy significantly different from the correlation r_2 between female smoking and female life expectancy?') we have to use a slightly different test. We convert r to a related number r' and work out a Z score from those:

$$r' = 0.5 \ln \left| \frac{1 + r}{1 - r} \right|$$

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Then we look up our value of z in a table of the standard normal distribution to get our p value.

'I have two values of r, but they are not independent; are they different?'

Suppose we measured the number of GCSE points acquired by a group of 16-year-olds, then measure the number of A-Level points acquired by the same people aged 18, then measure their annual income when they are 30. We could calculate a correlation between any two of these variables. We could also ask whether the correlation between GCSE scores and income was better / worse than the correlation between A-

Level scores and income. But these are clearly not independent correlations, because they were all based on the same people, and so there will probably be a correlation between GCSE scores and A-Level scores that we must take into account. If our three correlations are r_A, r_B, and r_C, and we want to know if the difference between r_A and r_B is significant, then the null hypothesis is that r_A and r_B are the same, and we can calculate a t statistic (with $n - 3$ df) like this:

$$|R| = (1 - r_{AB}^2 - r_{AC}^2 - r_{BC}^2) + 2 r_{AB} r_{AC} r_{BC}$$

$$t_{n-3} = (r_{AB} - r_{AC}) \sqrt{\frac{(n-1)(1+r_{BC})}{2\left(\frac{n-1}{n-3}\right)|R| + \frac{(r_{AB}+r_{AC})^2}{4}(1-r_{BC})^3}}$$

This is effectively a statistical test for partial correlations. The partial correlation coefficient will answer the question 'what is the correlation between X and Y, taking account of Z?' This test will answer the question 'is the correlation r_{xy} significantly different from the correlation r_{xz}, taking into account the fact that these two correlations are themselves related (non-independent)?'

'Is my value of r different from (a particular value)?'

Suppose we have a sample with a correlation of $r = 0.3$, and we want to know if this differs from a correlation of 0.5. The null hypothesis is that the sample $r = 0.3$ came from a population with $\rho = 0.5$. We can calculate a Z score like this:

$$r' = 0.5 \ln\left|\frac{1+r}{1-r}\right|$$

$$z = \frac{r' - \rho'}{\sqrt{\frac{1}{n-3}}}$$

If I calculate r, what are the confidence limits on ρ?

The expression for z above tells us that

$$\rho' = r' + \frac{z}{\sqrt{n-3}}$$

so we can calculate confidence intervals from appropriate critical values of z for a two-tailed α:

$$CI(\rho') = r' \pm z_{\alpha/2} \frac{1}{\sqrt{n-3}}$$

If you want 95% confidence intervals, $z_{\alpha/2}$ would be 1.96. Once you've worked out confidence intervals for ρ', we can convert them back to ρ to get our final answer:

$$\rho' = 0.5 \ln\left|\frac{1+\rho}{1-\rho}\right| \Rightarrow \rho = \frac{e^{2\rho'} - 1}{e^{2\rho'} + 1}$$

I have a group of subjects and have worked out a correlation for each one. Is this correlation significant for my whole group?

Stop and go back a stage. Suppose you have a group of 20 rats and you measure their performance on a test of attention and, simultaneously, the levels of the neurotransmitter acetylcholine in parts of their brain. Is there a relationship between acetylcholine and attentional performance? If you only make one measurement per rat, the problem is easy; you have 20 measurements of two variables, and can correlate them as usual. If you've made 100 measurements for *each* rat, the problem is harder. What can you do?

- You **must not** lump all the measurements together to give 2000 different {*x*, *y*} pairs — because these observations are definitely not equally independent, since subsets of observations are likely to be related by virtue of having come from the same rat.

- To ask whether subjects with high levels of acetylcholine have high levels of performance — a *between-subjects* question — you could take each rat's *mean* performance and *mean* acetylcholine and conduct a correlation as normal (n = 20). (And if you'd made 60 observations on some rats and 105 observations on others, it wouldn't matter, because you'd take the mean across all these subjects. If you really felt that it was worth placing more weight on data from subjects that you obtained measurements from, you could conduct a *weighted analysis*, weighting for the number of observations per subject.)

If different rats have very different levels of acetylcholine, then we could end up with something like our wild-boar-and-runner-bean effect — for example, you might find a negative correlation across the group (rats with lots of acetylcholine do worse than rats with less acetylcholine), even though if you looked for it, you might find a positive correlation *within* each rat (when any given rat has what is a high level of acetylcholine *for that rat*, it performs better). So…

- If we want to know whether changes in one variable (acetylcholine) are paralleled by changes in the other variable (performance) in the same subject, and that this is consistent across subjects — a *within-subjects* question using data from multiple subjects — we can estimate the relationship within subjects using a very general technique, called general linear modelling. This particular way of using a general linear model (GLM) is called multiple regression or analysis of covariance (ANCOVA). The GLM technique will handle even more complicated problems, such as when we have two groups of rats (a control group and one that has had part of their brain destroyed) and we want to know whether the relationship between acetylcholine and performance is different in the two groups. We will not cover these advanced techniques in the IB course.

I want to predict a variable on the basis of many other variables, not just one.

Then you need multiple regression, which we're not going to cover.

Correlation/regression in Excel — relevant functions (see Excel help for full details)

COVAR(…)	*Population* covariance (i.e. divide-by-n formula). So to calculate r using this number, you need to divide this by the product of the 'population SDs' of X and Y, calculated using STDEVP(…) — or multiply COVAR(…) by n and then divide it by n–1, before dividing the result by the product of the sample SDs, calculated using STDEV(…).
CORREL(…)	Calculates r.
RANK(…)	Don't use it — it gets the ranks wrong when there are ties.
Tools → Data Analysis → Covariance	Calculates sample covariances (i.e. divide-by-n–1 formula).
Tools → Data Analysis → Correlation	Calculates r.
Tools → Data Analysis → Regression	Calculates r, a, b, p.

Bibliography

Howell, D. C. (1997). *Statistical Methods for Psychology*. Fourth edition, Wadsworth, Belmont, California.