Lecture 16: Correlation
Notation and Equations

Notation:

ρ: Population correlation (Greek letter rho)

r: Sample correlation

$\hat{Y}$ (Y-hat): A prediction of a subject's Y score, usually based on knowing their X score

MS_{error}: The variance in Y that can't be explained by X, based on the error between the prediction $\hat{Y}$ and the actual outcome Y

Formula for correlation. The easiest way to compute a correlation is to use z-scores. First, convert all the X scores into z-scores labeled z_X, and convert all the Y scores into z-scores labeled z_Y. Then multiply z_X and z_Y together for each subject. Finally, take the average.

$$r = \frac{\sum(z_X \cdot z_Y)}{n-1} \qquad (1)$$

Notice that when we take the average, we divide by $n - 1$ even though there are n subjects. This is because of degrees of freedom. Even though it looks like we're adding n numbers, if we rewrote the formula in terms of the raw scores instead of z-scores, one of the summands would algebraically disappear.

The other thing to be careful about here is remembering what n is. For a correlation, n is the number of subjects. Each subject has an X score and a Y score, so there are n X scores and n Y scores. In other words, n is not the total number of X and Y scores combined; that's $2n$.

Fun facts about correlation. The value of a correlation tells us about both the direction of a relationship and its strength.

First, the sign of r tells whether the variables are positively or negatively related. If r is positive, that implies that the products $z_X \cdot z_Y$ in Equation 1 tend to be positive, meaning the positive z_Xs tend to go with positive z_Ys and the negative z_Xs tend to go with negative z_Ys. Because positive z-scores correspond to raw scores above the mean, this means that subjects with above-average X scores tend to have above-average Y scores. The same goes for below-average scores. Therefore the bigger X is, the bigger Y tends to be, so X and Y are positively related.

If r is negative, then the story is reversed. In this case, the products $z_X \cdot z_Y$ tend to be negative, meaning the positive z_Xs tend to go with negative z_Ys and vice versa. This means that subjects with above-average X scores tend to have below-average Y scores, and vice versa. Therefore, the bigger X is, the smaller Y tends to be, so X and Y are negatively related.

If r equals zero, then there's no linear relationship between the variables. This could mean that the variables are independent, i.e. knowing one of them doesn't give any information about the other. However, r could also be zero if the variables have a *nonlinear* relationship, such as one that first goes up and then goes down. Therefore we have to be

careful about interpreting a zero correlation. It's best to make a scatterplot of the data and visually check for a nonlinear relationship that the correlation couldn't detect.

Putting all this together, we have the following simple rule for interpreting the direction of a correlation.

$$\begin{aligned} r > 0 &\rightarrow \text{positive relationship} \\ r < 0 &\rightarrow \text{negative relationship} \\ r = 0 &\rightarrow \text{no } \underline{\text{linear}} \text{ relationship} \end{aligned} \qquad (2)$$

The other thing that correlation tells us is the strength of a relationship. The strength is indicated by the absolute value of *r*, i.e. how large *r* is without regard for whether it's positive or negative. The further *r* is from zero, the stronger the relationship. For example, $r = .5$ and $r = -.5$ indicate equally strong relationships; the only difference is in their directions. A correlation near zero means a weak relationship, a correlation closer to +1 or -1 means a stronger relationship, and a correlation of exactly ±1 means a perfect relationship. It's important to remember that the correlation can never be more extreme than ±1. In other words *r* is always between -1 and +1. For example, a correlation of +2 or -2 is mathematically impossible.

If you're interested, here's one way to see why $|r|$ can never be bigger than 1. The strongest possible relationship between two variables arises when they're identical. In that case, we'd be computing the correlation between *X* and *X*, and Equation 1 reduces to $r = \Sigma(z_X)^2/(n\text{-}1)$. This is the formula for the variance of *z* (there's no mean in this formula, but the mean of z-scores is always 0). By definition, z-scores have a variance of 1, so that's our answer: $r = 1$. In conclusion, the strongest possible relationship between variables leads to a correlation of 1. We could do the same thing with negative relationships, by computing the correlation between *X* and –*X*, and we'd end up with $r = -\text{var}(z_X)$, or $r = -1$. All of this comes from the fact that we use standardized scores (z-scores) to compute correlation. Using z-scores effectively standardizes the correlation, putting it in a range from -1 to 1.

<u>Mean squared error</u>. When we use *X* to predict *Y*, we call the prediction $\hat{Y}$. The error of our prediction is the difference $Y - \hat{Y}$. We want to keep this error as close to zero as possible, and we do this by minimizing the square of the error (which is always positive), averaged over all of the data. This leads to the formula for mean squared error. Later on, we'll also call this the *residual* mean square, because it's the variability that's left over after we use *X* to predict *Y*.

$$MS_{\text{error}} = \frac{\sum\left(Y - \hat{Y}\right)^2}{n-1} \qquad (3)$$

<u>Correlation and Prediction</u>. Another way to think of the relationship between two variables is to ask how well we can use one to predict the other. There are infinitely many ways to use one variable to predict another, but here we're just interested in linear prediction. This basically means drawing a straight line through your scatterplot, so that

the line is as close as possible to the data. For any value of X, the height of the line gives $\hat{Y}$, which is the predicted value of Y.

The "best" prediction line is the one that minimizes MS_{Error}. Directly working out what this line is requires calculus (ask me and I'll be happy to show you), but fortunately there's an easy shortcut that uses the correlation. If we plot the data using z-scores instead of raw scores, meaning we make a scatterplot of z_X versus z_Y, finding the best prediction line is simple. It turns out to be the line going through the origin (0,0) with a slope of r.

$$z_{\hat{Y}} = r \cdot z_X \tag{4}$$

If r is positive, then the prediction line slopes upward, because X and Y are positively related. The opposite is true if r is negative. If $r = 0$, then the prediction line is flat, because X tells us nothing about Y so the prediction is the same regardless of what X is. More generally, the stronger the correlation is, the steeper the prediction line is (either up or down, depending on the sign of r), because knowing X tells us more about Y.

The other connection between correlation and prediction is that the correlation tells us how well one variable predicts the other. That is, we can use r to find out the mean squared error of our best prediction (Eq. 4), which turns out to be

$$MS_{\text{error}} = \left(1 - r^2\right) \cdot s_Y^2 \tag{5}$$

This quantity, $(1 - r^2) \cdot s_Y^2$, is called the residual variance, because it's the variance in Y that's left over after we use the information from X. To get a measure of how well X can predict Y, we want to know the amount of variance in Y that X can "explain," meaning how much the variance goes down when we use X to improve our prediction. The explained variance equals the original variance, s_Y, minus the residual variance. Using the equation above for residual variance, the formula for explained variance works out to be pretty simple.

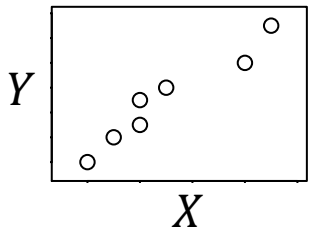
$$\text{explained variance} = r^2 \cdot s_Y^2 \tag{6}$$

In other words, r^2 tells us the fraction or proportion of the variance in one variable that we can eliminate by using the other variable as a predictor. We get the same fraction if we use X to predict Y or Y to predict X (because the correlation is the same either way). Notice that the proportion of explained variance doesn't depend on whether r is positive or negative. Because the proportion equals r^2, it only depends on the strength of the correlation, not the direction. If r equals ±1, the two variables are perfectly related, meaning we can use one to predict the other exactly. In other words, X explains everything about Y (and vice versa), which fits with the fact that $r^2 = 1$. If r equals zero, knowing one variable does us no good in predicting the other (at least if we limit ourselves to linear prediction), and the explained variance is zero. Finally, if the correlation is some intermediate value, then we can use one variable to explain some, but not all, of the variance in the other variable. For example, if r is +.5 or -.5, then X can explain 25% of the variance in Y (and vice versa).

In summary, r gives us two ways of thinking about the strength of the relationship between two variables. First, we can think of r as the correlation, meaning a measure of how well the data lie along a straight line. The further r is from zero, and the closer it is to ±1, the stronger the relationship. Second, we can think of r^2 as a measure of how well one variable

can explain the other, meaning the proportion of the variance that goes away when we use one variable to predict the value of the other.

Example of calculating correlation. Assume we have samples for two variables, X = [5, 8, 3, 9, 2, 4, 4] and Y = [8, 10, 4, 13, 2, 7, 5]. The samples come from the same set of subjects, meaning the first subject has scores of $X = 5$ and $Y = 9$, and so on. Here's a scatterplot of the data. Each point corresponds to one subject, with its horizontal position indicating that subject's X score and its vertical position indicating that subject's Y score.



You can see from the plot that the variables have a strong positive relationship. For example, the subject with the highest X score (the 4th subject in the sample) also has the highest Y score. This subject is shown by the point in the upper right of the scatterplot. The subject shown by the point in the lower left has the lowest X score and also the lowest Y score.

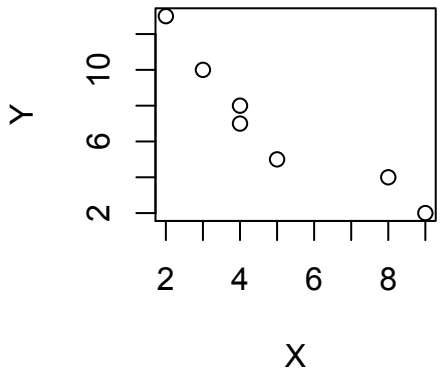
Now let's compute the correlation between X and Y, using Equation 1. The first step is to convert the raw scores to z-scores. For X, we subtract the mean of the X sample from each individual score and then divide by the sample standard deviation of X. This gives us $z_X = (X - M_X)/s_X$. We do the same with Y, making sure now to use the mean and standard deviation of the Y sample, to get $z_Y = (Y - M_Y)/s_Y$. The next step is to multiply the z-scores together. For each subject, we multiply their z-score for X and their z-score for Y to get $z_X \cdot z_Y$. Finally, we take the average by adding up all the $z_X \cdot z_Y$ values and dividing by $n - 1$. The following table goes through this whole process.

X	Y	z_X	z_Y	$z_X \cdot z_Y$
5	8	.00	.27	.00
8	10	1.16	.80	.93
3	4	-.77	-.80	.62
9	13	1.55	1.60	2.48
2	2	-1.16	-1.34	1.55
4	7	-.39	.00	.00
4	5	-.39	-.53	.21
$M_X = 5$	$M_Y = 7$			$\sum(z_X \cdot z_Y) = 5.80$
$s_X = 2.6$	$s_Y = 3.7$			$r = \sum(z_X \cdot z_Y)/(n\text{-}1) = .97$

Now think about how the positive relationship between X and Y leads to a positive value of r. The subjects who have above-average values of X (the 2nd and 4th subjects) also have above-average values of Y. Therefore these subjects have positive values of both z_X and z_Y, which leads $z_X \cdot z_Y$ to be positive as well. The subjects with below-average X scores also have

below-average Y scores. Therefore z_X and z_Y are both negative for these subjects, and so once again $z_X \cdot z_Y$ is positive. In other words, the correspondence between X and Y leads $z_X \cdot z_Y$ to be consistently positive (or zero) for every subject. This leads r to be both positive and large.

As a second example, imagine the scores are X = [5, 8, 3, 9, 2, 4, 4] and Y = [5, 4, 10, 2, 13, 7, 8]. These are the same numbers as before, but rearranged so that they pair up differently. The scatterplot shows a negative relationship in the new data.



The table below calculates the correlation for the new data. This time, the subjects with below-average X scores have above-average Y scores, and vice versa. Therefore, every subject has one negative z-score and one positive z-score (except for z-scores of zero). As a consequence, $z_X \cdot z_Y$ is negative (or zero) for every subject, which leads r to be strongly negative.

X	Y	z_X	z_Y	$z_X \cdot z_Y$
5	5	.00	-.53	.00
8	4	1.16	-.80	-.93
3	10	-.77	.80	-.62
9	2	1.55	-1.33	-2.07
2	13	-1.16	1.60	-1.86
4	7	-.39	.00	.00
4	8	-.39	.27	-.10
M_X = 5	M_Y = 7			$\sum(z_X \cdot z_Y)$ = -5.59
s_X = 2.6	s_Y = 3.7			$r = \sum(z_X \cdot z_Y)/(n\text{-}1)$ = -.93