Final Remarks About Correlation

Partial correlation. For the last few lectures, we have been discussing correlation as a measure of the relationship between two variables. But the relationship between any two variables X and Y could be influenced by their mutual associations with a third variable Z. Suppose that three variables have a multivariate normal distribution,

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \\ \mu_z \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} \right).$$

The simple (marginal) correlation between X and Y is

$$\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\,\sigma_{yy}}} = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y}.$$

Using properties of the multivariate normal distribution (Lecture 5), we can also find the correlation between X and Y at any fixed value of Z. The conditional distribution of $(X,Y)^T$ given Z is bivariate normal with covariance matrix

$$\begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} - \begin{pmatrix} \sigma_{xz} \\ \sigma_{yz} \end{pmatrix} (\sigma_{zz})^{-1} \begin{pmatrix} \sigma_{xz} & \sigma_{yz} \end{pmatrix},$$

which becomes

$$\begin{pmatrix} \sigma_{xx\cdot z} & \sigma_{xy\cdot z} \\ \sigma_{yx\cdot z} & \sigma_{yy\cdot z} \end{pmatrix} = \begin{pmatrix} \sigma_{xx} - \sigma_{xz}^2/\sigma_{zz} & \sigma_{xy} - \sigma_{xz}\sigma_{yz}/\sigma_{zz} \\ \sigma_{xy} - \sigma_{xz}\sigma_{yz}/\sigma_{zz} & \sigma_{yy} - \sigma_{yz}^2/\sigma_{zz} \end{pmatrix}.$$

The correlation between X and Y at a fixed value of Z is

$$\rho_{xy\cdot z} = \frac{\sigma_{xy\cdot z}}{\sqrt{\sigma_{xx\cdot z}\,\sigma_{yy\cdot z}}}.$$

With a little algebraic manipulation, we can write it in terms of the simple correlations,

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

This quantity is called the "partial correlation" between X and Y given Z, although a more appropriate name might be "conditional correlation." It measures the relationship between X and Y after accounting for their mutual associations with Z. If $\rho_{xy\cdot z} = 0$, then we can say that the association between X and Y is fully explained by Z.

Notice that the partial correlation between X and Y given Z does not depend on Z. Under the multivariate normal model, the correlation between any two variables given a third is the same for all values of the third. This is another aspect of the "homoscedasticity" property of the multivariate normal distribution. The multivariate normal model has no interactions. (An interaction is a relationship among three variables in which the correlation between two of them changes in relation to the third.)

Just by examining this formula, we see two interesting facts.

1. $\rho_{xy\cdot z} = \rho_{xy}$ if $\rho_{xz} = \rho_{yz} = 0$.
2. $\rho_{xy\cdot z} = 0$ if $\rho_{xy} = \rho_{xz}\rho_{yz}$.

Property 1 says that the simple correlation and the partial correlation are the same when the variable we're conditioning on is independent of both of the variables in question. Property 2 gives us the size of a spurious correlation between two variables when a third variable is ignored. Suppose that the relationship between X and Y is fully explained by Z ($\rho_{xy\cdot z} = 0$). If we ignore Z, then the two variables will appear to be related, and the simple correlation between them (ρ_{xy}) will be equal to the product of their simple correlations with the third variable ($\rho_{xz}\rho_{yz}$). If $\rho_{xy} \neq \rho_{xz}\rho_{yz}$, then the relationship between X and Y is not fully eXplained by Z.

Recall that ρ^2_{xy}, the squared simple correlation between X and Y, can be interpreted as the proportion of the variance in Y explained by X (or vice-versa). A similar argument can be used to interpret $\rho^2_{xy\cdot z}$. The variance in

Y not explained by Z is

$$V(Y \mid Z) \;=\; \sigma_{yy\cdot z} \;=\; \sigma_{yy}(1-\rho_{yz}^2).$$

The variance in Y not explained by X and Z is

$$V(Y \mid X, Z) \;=\; \sigma_{yy\cdot xz}.$$

With a little algebra, one can show that

$$\sigma_{yy\cdot xz} \;=\; \sigma_{yy\cdot z}(1-\rho_{xy\cdot z}^2).$$

Therefore, $\rho_{xy\cdot z}^2$ is **the proportion of variance in Y explained by X after accounting for the variance explained by Z.** Exchanging the roles of Y and X, we can also show that

$$\sigma_{xx\cdot yz} \;=\; \sigma_{xx\cdot z}(1-\rho_{xy\cdot z}^2),$$

so $\rho_{xy\cdot z}^2$ is also **the proportion of variance in X explained by Y after accounting for the variance explained by Z.**

Discrepancies between simple and partial correlations. We defined the partial correlation between X and Y given Z as

$$\rho_{xy\cdot z} \;=\; \frac{\rho_{xy} \;-\; \rho_{xz}\rho_{yz}}{\sqrt{(1-\rho_{xz}^2)\,(1-\rho_{yz}^2)}}. \qquad (1)$$

The simple correlation between X and Y, which we write as ρ_{xy}, tells us little or nothing about $\rho_{xy\cdot z}$. The partial correlation may be larger or smaller than the simple correlation. They may even have opposite signs. (With categorical variables, that phenomenon is called "Simpson's paradox.")

Suppose, for example, that you collect two variables,

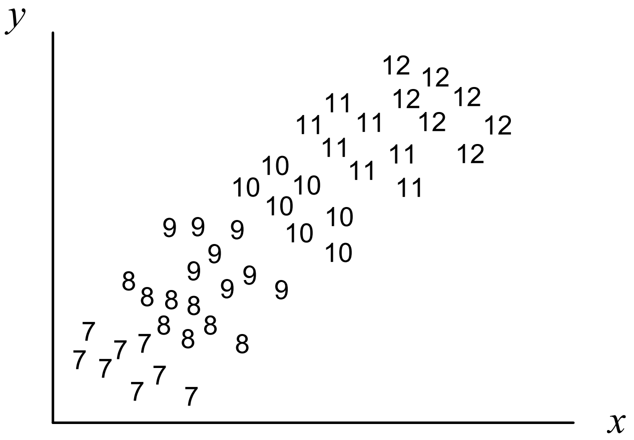
$$\begin{aligned} X &= \text{score on a math achievement test, and} \\ Y &= \text{number of alcoholic beverages} \\ &\quad\ \text{consumed in the last 30 days} \end{aligned}$$

for a representative sample of American middle and high school students. It's quite possible that the correlation between these two variables will be *positive*. Students with higher math achievement scores will tend to have higher levels of alcohol consumption. Does this mean that math achievement is a risk factor for adolescent alcohol use? Of course not. Common sense tells us that the opposite should be true; higher levels of math achievement are probably associated with reduced levels of alcohol use.

The problem with the simple correlation is that it does not control for one obvious confounder: age. Suppose that, in addition to X and Y, we also record

$$Z = \text{grade in school } (7, 8, \ldots, 12).$$

Within any grade, we are likely to find a significant negative correlation between X and Y. Overall, however, the correlation between X and Y may be positive because both are positively correlated with grade. Here is an example of how the data might look, with the effects greatly exaggerated.



Inferences about partial correlation. We may estimate the partial correlation by replacing each population correlation in the formula (1) by its corresponding sample correlation,

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

For example, suppose we examine a sample of $n = 100$ children and record these three variables:

$$\begin{aligned} X &= \text{reading comprehension score} \\ Y &= \text{body weight} \\ Z &= \text{age} \end{aligned}$$

Suppose the sample correlation matrix looks like this.

$$\begin{pmatrix} r_{xx} & r_{xy} & r_{xz} \\ r_{yx} & r_{yy} & r_{yz} \\ r_{zx} & r_{zy} & r_{zz} \end{pmatrix} = \begin{pmatrix} 1.000 & .616 & .827 \\ .616 & 1.000 & .732 \\ .827 & .732 & 1.000 \end{pmatrix}.$$

The high correlation between reading score and age ($r_{xz} = .827$) is to be expected, and so is the high correlation between weight and age ($r_{yz} = .732$). But what about the correlation between reading score and weight ($r_{xy} = .616$)? Does this suggest that children's reading will tend to improve if they gain weight? Notice that

$$r_{xz} \times r_{yz} \; = \; .827 \times .732 \; = \; .605,$$

which is very close to $r_{xy} = .616$. This suggests that the relationship between X and Y might be explained by their mutual associations with Z. The estimated partial correlation between X and Y given Z is

$$r_{xy\cdot z} \; = \; \frac{.616 - (.827)(.732)}{\sqrt{(1 - .827^2)\,(1 - .732^2)}} \; = \; .028,$$

which is very close to zero.

Is this value significantly different from zero? The theory for the distribution of the partial correlation coefficient is remarkably similar to that for an ordinary correlation. Testing the null hypothesis $H_0 : \rho_{xy\cdot z} = 0$ is equivalent to testing $H_0 : \beta_1 = 0$ in the normal linear regression model

$$Y_i \mid X_i, Z_i \;\sim\; N(\,\beta_0 + \beta_1 X_i + \beta_2 Z_i,\, \sigma_{yy\cdot xz}\,).$$

(This linear regression model will hold if $(X, Y, Z)^T$ has a multivariate normal distribution. But it is more general, because it may also cover situations where X or Z are not normal.) The test of $H_0 : \beta_1 = 0$ is based on the t-statistic

$$t \;=\; \hat{\beta}_1 / SE(\hat{\beta}_1),$$

where $\hat{\beta}_1$ is the ordinary least-squares estimate of β_1, and $SE(\hat{\beta}_1)$ is its standard error. This statistic is compared to a Student's t-distribution with $n-3$ degrees of freedom. (For a simple correlation, the degrees of freedom were $n-2$. They are now $n-3$, because we have estimated one additional parameter, β_2.) The relationship between this t-statistic and $r_{xy\cdot z}$ is the same as before,

$$r^2_{xy\cdot z} \;=\; \frac{t^2}{t^2 + df},$$

except that now $df = n-3$. Solving for t gives

$$t \;=\; \mathrm{sign}(r_{xy\cdot z}) \sqrt{\left(\frac{r^2_{xy\cdot z}}{1 - r^2_{xy\cdot z}} \right) df}.$$

Applying this to our hypothetical example, we found that $r_{xy\cdot z} = .028$ in a sample of $n = 100$. Converting this to a

t-statistic, we get

$$t \;=\; \sqrt{\left(\frac{.028^2}{1-.028^2}\right)\times 97} \;=\; 0.276,$$

which is definitely not significant at the .05 level.

The method for constructing a confidence interval for a partial correlation is also remarkably similar to that of a simple correlation. For a simple correlation, we used Fisher's approximation

$$g(r_{xy}) \;\sim\; N\left(\, g(\rho_{xy}),\, \frac{1}{n-3}\,\right),$$

where $g(r) = \tanh^{-1}(r)$. For a partial correlation, the approximation is

$$g(r_{xy\cdot z}) \;\sim\; N\left(\, g(\rho_{xy\cdot z}),\, \frac{1}{n-4}\,\right).$$

An approximate confidence interval for $Z = g(\rho_{xy\cdot z})$ goes from

$$Z_1 \;=\; g(r_{xy\cdot z}) - 1.96\sqrt{\frac{1}{n-4}}$$

to

$$Z_2 \;=\; g(r_{xy\cdot z}) + 1.96\sqrt{\frac{1}{n-4}},$$

and the corresponding interval for $\rho_{xy\cdot z}$ goes from $\tanh(Z_1)$ to $\tanh(Z_2)$. This is how you might do it in R.

```
> rxy <- cor(x,y)
> rxz <- cor(x,z)
> ryz <- cor(y,z)
> rxy.z <- ( rxy - rxz*ryz ) / sqrt( ( 1-rxz^2 ) * ( 1-ryz^2 ) )
> z.low  <- atanh(rxy.z) - 1.96 * sqrt( 1 / (n-4) )
> z.high <- atanh(rxy.z) + 1.96 * sqrt( 1 / (n-4) )
> r.low  <- tanh(z.low)
> r.high <- tanh(z.high)
```

Partial correlation and residuals. Another way to compute the sample partial correlation $r_{xy\cdot z}$ is

- regress X on Z and save the residuals;
- regress Y on Z and save the residuals; then
- compute the simple correlation between the two sets of residuals.

If you have already taken a course that involves regression, you know what the residuals are. If you haven't, don't worry; we will learn about residuals very soon. Here is how you would do it in R.

```
> result <- lm( x ~ z )
> res.1 <- result$residuals
> result <- lm( y ~ z )
> res.2 <- result$residuals
> rxy.z <- cor( res.1, res.2 )
```

The value of $r_{xy\cdot z}$ that you would get from this procedure is identical to what you would get from the formula (2).

Partial correlation given additional variables. The formula for partial correlation (1) also holds for conditioning on additional variables. For example, the partial correlation between X and Y given Z and W can be written in terms of partial correlations given only W,

$$\rho_{xy\cdot wz} \;=\; \frac{\rho_{xy\cdot w} \;-\; \rho_{xz\cdot w}\rho_{yz\cdot w}}{\sqrt{\left(1-\rho^2_{xz\cdot w}\right)\left(1-\rho^2_{yz\cdot w}\right)}}.$$

By repeatedly applying this principle, we can recursively compute partial correlations given any number of variables from the simple correlations.

In practice, however, statisticians do not compute partial correlations this way. Rather, they tend to compute partial correlations by fitting regression models and transforming the t-statistics to correlations by the formula

$$r \;=\; \mathrm{sign}(t)\sqrt{\frac{t^2}{t^2+\mathit{df}}},$$

where df is the sample size n minus the number of coefficients in the regression model. The resulting r is the partial correlation between the response variable and the predictor in question, given all the other predictors in the model. For example, suppose we fit the linear regression model

$$Y \;=\; \beta_0 \;+\; \beta_1 X \;+\; \beta_2 W \;+\; \beta_3 Z \;+\; \epsilon.$$

If we take the t-statistic for β_1, which is $t = \hat{\beta}_1/\mathrm{SE}(\hat{\beta}_1)$, and convert it to r using the formula above, the result value will be $r_{xy\cdot wz}$, the estimated partial correlation between X and Y given W and Z. (In this case, the degrees of freedom would be $\mathit{df} = n-4$, because the regression model has four unknown β's.)

Yet another way to compute the partial correlation between X and Y given any set of variables is to

- regress X on that set of variables and save the residuals;
- regress Y on that set of variables and save the residuals; then
- compute the simple correlation between the two sets of residuals.

Partial correlation given a grouping variable. Sometimes we want to estimate the partial correlation between two variables, X and Y, given a non-numeric grouping variable. For example, suppose you collected the answers to two questions from a sample of likely voters:

X	=	In the next presidential election, how likely are you to vote for the Republican candidate? (1=very unlikely, ..., 5=very likely)
Y	=	What was your income last year?

As we pointed out in the last lecture, the correlation between these two variables in your sample is likely to be positive. If you computed the average values of X and Y within geographic areas, however, the correlation between these average scores would be negative. The discrepancy would arise because both of these variables are related to geography. People who live in the same area tend to have similar incomes, and people who live in the same area also tend to have similar political leanings.

Suppose we wanted to answer the question, "What portion of the association between X and Y is not explained by geography?" If Z = geographic area was a numeric variable, we could address this question by computing $r_{xy \cdot z}$ as described above. But in this case, Z consists of nominal categories. The partial correlation between X and Y given these categories can be computed as follows.

- Subtract from each subject's value of X the average value of X for his or her geographic area.
- Subtract from each subject's value of Y the average value of Y for his or her geographic area.
- Compute the simple correlation between these de-meaned values of X and Y.

This is how you might do it in R.

```
> n <- length(x)
> x.new <- numeric(n)  # create a blank vector to hold the de-meaned x
> y.new <- numeric(n)  # create a blank vector to hold the de-meaned y
> for( k in unique(z) ){
+    x.new[ z==k ] <- x[ z==k ] - mean( x[ z==k ] )
+    y.new[ z==k ] <- y[ z==k ] - mean( y[ z==k ] )}
> rxy.z <- cor(x.new, y.new)
```

Another way to compute $r_{xy \cdot z}$ is to regress X and Y on a set of dummy variables defined by the levels of Z, and then correlate the residuals. We will learn about this soon.

Introduction to regression. Regression is different from correlation. Correlation is symmetric in the following sense: The correlation between X and Y is the same as the correlation between Y and X. In simple linear regression, however, we designate one of the variables (Y) to be a response and then predict it from one or more X's. Correlation describes the joint distribution of X and Y, whereas regression describes the conditional distribution of Y given X.

In regression analysis, we treat the X-variables (i.e., the predictors) as fixed constants. In many applications, the X's are random variables not under the control of the investigators. Treating X as fixed is merely a technique to avoid specifying a model for the joint distribution of the predictors, which is usually regarded as a nuisance, to focus attention on the question of interest—namely, the "effect" of the predictors on the response. (Here we use the term "effect" very loosely. It does not mean that the relationship between X and Y is causal. The meaning of causality and techniques for causal inference will be discussed later in the semester.)

A typical linear regression model has the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon, \tag{3}$$

where ϵ is a normally distributed random error with mean zero and variance σ^2. The unknown parameters are $\beta_0, \ldots, \beta_p$ and σ^2. For notational reasons, it will sometimes be more convenient to write the model as

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon. \tag{4}$$

When written this way, it is understood that in most cases the first predictor variable is actually a constant ($X_1 \equiv 1$). Using the latter notation, we can collect the predictors and coefficients into vectors,

$$\begin{aligned} X &= (X_1, X_2, \ldots, X_p)^T, \\ \beta &= (\beta_1, \beta_2, \ldots, \beta_p)^T, \end{aligned}$$

and write the model as

$$Y \sim N(\,X^T\beta,\, \sigma^2),$$

where conditioning on X is understood.

To derive methods of *exact* inference (exact tests and confidence intervals) about the β's, we will need to assume that the error term ϵ is normally distributed. Although this seems to limit the usefulness and generality of regression, it turns out that this assumption is not crucial for most purposes. Many (but not all) of the procedures that we will cover are not sensitive to moderate departures from normality, especially if the sample size is large. It is fair to say that, in linear regression, the assumption of normality is the least important assumption that we will make. The other assumptions, which are more crucial, are

- the constancy of the error variance σ^2, and
- the independence of the errors ϵ across units.

If the X's are actually random variables, then we are also implicitly assuming that the ϵ's are uncorrelated with the X's. That is, we assume that ϵ has mean zero for all units regardless of their values of X.

The no-predictors model. Most textbooks on regression begin with a single predictor,

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{5}$$

which is called "simple linear regression," and then move up to the model (3), which is called "multiple linear regression." In this course, however, we will start with

$$Y = \beta + \epsilon, \tag{6}$$

which can be regarded as a special case of (4) with $p = 1$ and $X_1 \equiv 1$. This model, which we call the "no-predictors" model, is ridiculously simple. But the primary results and the intuition we develop about it will immediately generalize to the more complicated settings.

If the error ϵ is assumed to be normally distributed with mean 0 and variance σ^2, then (6) is equivalent to $Y \sim N(\beta, \sigma^2)$. The data to fit this model will consist of n independent observations of Y, which we denote by $Y_1, Y_2, \ldots, Y_n$. If we collect these into a vector,

$$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T,$$

then the no-predictors model can be written in multivariate form,

$$\boldsymbol{Y} \sim N(\mu, \sigma^2 I\,), \tag{7}$$

where $\mu = \beta \cdot 1$, and $1 = (1, 1, \ldots, 1)^T$. Although the mean of $\boldsymbol{Y}$ is μ, which is a vector of length n, it does not contain n unknown parameters. We have constrained μ to lie within $\mathcal{R}(1)$, the linear space spanned by the vector 1, so μ is a function of the single parameter β.

The variance σ^2 is often considered to be a nuisance parameter. Although it may not be of direct interest, we need to pay attention to it, because it will affect the precision of the estimate of β and the prediction of future observations of Y.