The Bottom Line

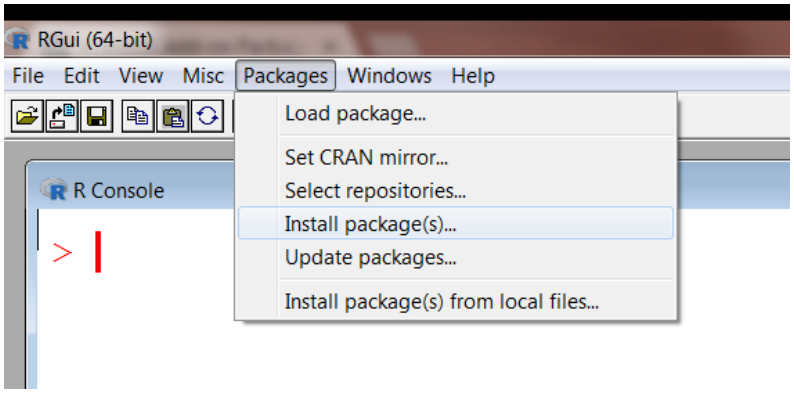
In the previous lecture, we dashed across the cafeteria, knocking chairs and tables over in a rush to look at partial correlation examples. In this lecture we take a few moments to develop the idea of partial correlation more carefully.

For those of us eager to get the "bottom line", we start right out by saying that we can think of a time series as a set of (presumably related!) random variables, $X_1,\ X_2, \ldots X_t, \ldots$ and start talking about *conditional dependencies.*

Working to Understand the Math: A Story with Pictures

It often happens that we have several related variables, and the correlations between them have some redundancy. I usually think of this as "Hey, you already told me that, please tell me something new!", but that's not very mathematical. Perhaps better stated, you would like to control for the presence of terms already in a model when deciding whether a proposed variable is useful. Without digressing too far we can say that we'd like to measure a correlation between some random variables of interest with the effect of other variables removed. In particular, when assessing whether we wish to include the p^{th} term, can we measure the correlation at lag p that has not been accounted for in a model with p-1 coefficients?

Here's a quick regression example. A data set that you will see in various textbooks is also available in the *isdals* library. If it's not on your system you may have to do the typical installation process.



Once installed, you can access a data set called *bodyfat* by calling

library(isdals)

data(bodyfat)

The description is as follows:

> *It is expensive and cumbersome to determine the body fat in humans as it involves immersion of the person in water. This dataset provides information on body fat, triceps skinfold thickness, thigh circumference, and mid-arm circumference for twenty healthy females aged 20 to 34. It is desirable if a model could provide reliable predictions of the amount of body fat, since the measurements needed for the predictor variables are easy to obtain.*

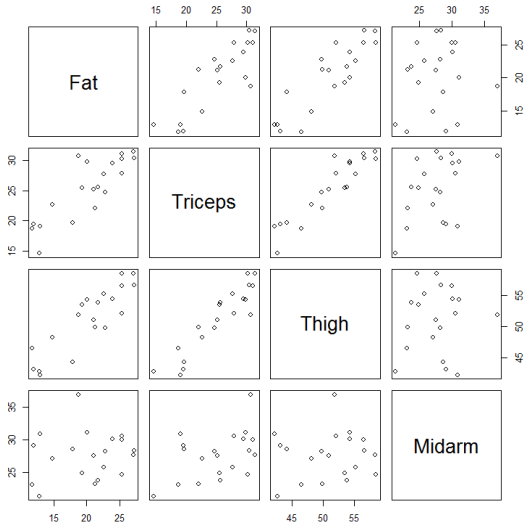
The variables are

`Fat:`	body fat	`Triceps:`	triceps skinfold measurement
`Thigh:`	thigh circumference	`Midarm:`	mid-arm circumference

It probably won't shock you that these variables are related. We can see this with a pairs plot.

attach(bodyfat)

pairs(cbind(Fat, Triceps, Thigh, Midarm))



Clearly, *Fat* and *Triceps* (skin fold thickness) are highly correlated, *r=0.8432654*. But, so are *Fat* and *Thigh* circumference, *r= 0.8780896*.

```
cor( cbind( Fat, Triceps, Thigh, Midarm) )
```

	Fat	Triceps	Thigh	Midarm
Fat	1.0000000	0.8432654	0.8780896	0.1424440
Triceps	0.8432654	1.0000000	0.9238425	0.4577772
Thigh	0.8780896	0.9238425	1.0000000	0.0846675
Midarm	0.1424440	0.4577772	0.0846675	1.0000000

Since *Triceps* and *Thigh* are also clearly related *r= 0.9238425*, we wonder if we can measure the correlation of *Fat* and *Triceps*, after controlling for or "partialling out" *Thigh*. We first try to account for the effect of *Thigh* on both *Fat* and *Triceps* by regressing them on *Thigh*. After we remove the contribution of *Thigh*, we then find the correlation of *Fat* and *Triceps*. This is pretty easy to do; just use the *lm()* command we've previously discussed.

```
Fat.hat      = predict(lm(Fat~Thigh))
Triceps.hat  = predict( lm(Triceps~Thigh) )
cor( (Fat- Fat.hat), (Triceps- Triceps.hat) )        #returns 0.1749822
```

So, a great deal of the correlation between *Fat* and *Triceps* is accounted for by controlling for *Thigh* circumference. What happens when we control for both *Thigh* and *Midarm*? We do the calculation ourselves, and confirm by using the *pcor()* routine.

```
Fat.hat      = predict(lm(Fat~Thigh+Midarm))
Triceps.hat  = predict( lm(Triceps~Thigh+Midarm) )
cor( (Fat- Fat.hat), (Triceps- Triceps.hat) )        #returns      0.33815
pcor( cbind( Fat, Triceps, Thigh, Midarm) )
```

	Fat	Triceps	Thigh	Midarm
Fat	1.0000000	0.3381500	-0.2665991	-0.3240520
Triceps	0.3381500	1.0000000	0.9963725	0.9955918
Thigh	-0.2665991	0.9963725	1.0000000	-0.9926612
Midarm	-0.3240520	0.9955918	-0.9926612	1.0000000

Now, given a time series, we hope not to look all the way back to the first observation in thinking about our dependencies, especially since as time goes on this is likely to be less and less important. But, let's at least look back several observations, for example in an *AR(p)* process take a look back by *k* observations. Recall our model.

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

Now start writing partial autocorrelations at various lag spacings, These are just (sequentially ordered) random variables after all:

$$lag\ spacing\ 1: \quad \rho[X_t, X_{t-1}] = \frac{COV[X_t, X_{t-1}]}{V[X_t] \cdot V[X_{t-1}]} = (plain\ old)\ \ \rho(1)$$

$$lag\ spacing\ 2: \quad \rho[X_t, X_{t-2}] = \frac{COV[X_t,\ X_{t-2} \mid X_{t-1}]}{V[X_t | X_{t-1}] \cdot V[X_{t-2} | X_{t-1}]}$$

$$lag\ spacing\ 3: \quad \rho[X_t, X_{t-3}] = \frac{COV[X_t,\ X_{t-3} \mid X_{t-1}, X_{t-2}]}{V[X_t | X_{t-1}, X_{t-2}] V[X_{t-3} | X_{t-1}, X_{t-2}]}$$

$$lag\ spacing\ k: \quad \rho[X_t, X_{t-k}] = \frac{COV\left[X_t,\ X_{t-k} \mid X_{t-1}, X_{t-2}, X_{t-(k-1)}\right]}{V\left[X_t | X_{t-1}, X_{t-2}, X_{t-(k-1)}\right] V\left[X_{t-k} | X_{t-1}, X_{t-2}, X_{t-(k-1)}\right]}$$

If you are unfamiliar with this notation, we will explain it below[1]. And, of course I snuck a few more known outcomes in at lag spacings past 2, but this should be intuitive enough, at least if you have seen these types of concepts before.

This all probably looks a bit cumbersome. And, it's not obvious how we would estimate these quantities (except, perhaps, in the multivariate normal case). The basic idea is to find the correlation between random variables X_t and X_{t-k} after accounting for or "partialling out" the linear effects of the intervening random variables.

[1] Please see the section appropriately called "Digging into the Notation"

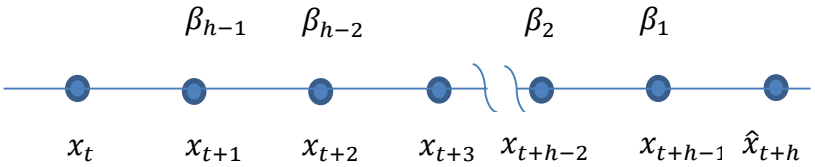
Making This Happen for a Time Series

There is a nice treatment of this material in (Shumway & Stoffer, 2013). They denote the regression of a term x_{t+h} on preceeding terms $x_{t+h-1}, x_{t+h-2}, \ldots, x_{t+1}$ as $\hat{x}_{t+h}$.

That is,

$$\hat{x}_{t+h} = \beta_1 x_{t+h-1} + \beta_2 x_{t+h-2} + \cdots + \beta_{h-1} x_{t+1}$$

A picture might help. We estimate x_{t+h} as $\hat{x}_{t+h}$ by looking ***backward*** over the last several terms with

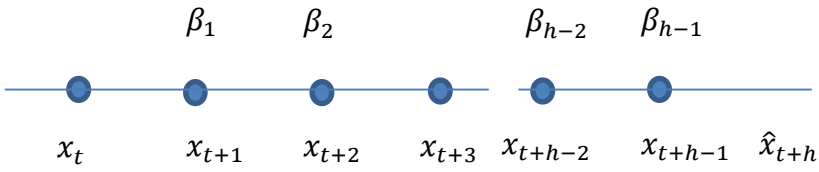


The subscripts on the β coefficients tell you how far away you are from the target variable.

Create the same type of representation looking in the other direction and estimate x_t, denoted $\hat{x}_t$, by looking ***forward*** over the next several terms. (We can use the same β values due to stationarity.)

$$\hat{x}_t = \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots + \beta_{h-1} x_{t+h-1}$$

Our picture now will be the following:



In the same spirit as the discussion with body fat, we define a partial autocorrelation function

$$\phi_{hh} \equiv corr[x_{t+h} - \hat{x}_{t+h},\ x_t - \hat{x}_t]$$

We remove the linear effects of all the terms between the two random variables we are focusing on as

$$corr[x_{t+h} - \hat{x}_{t+h},\ x_t - \hat{x}_t]$$

This is the partial correlation coefficient.

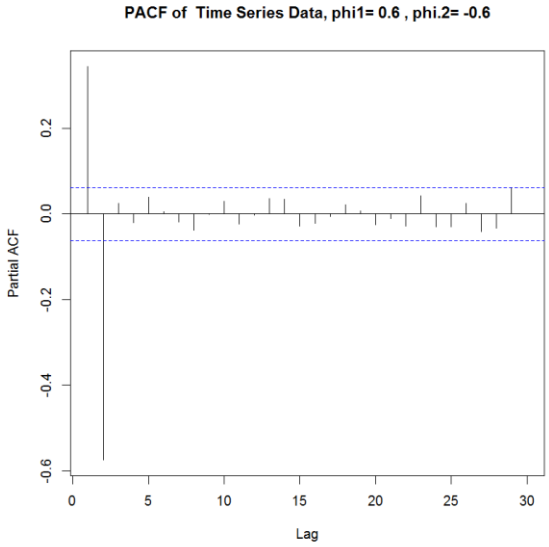
We can calculate (estimate, really) these quantities from a given time series and have another plot to use in our quest to understand the stochastic process that generated the data at hand. We will plot the ***P***artial ***A***uto ***C***orrelation ***F***unction (PACF). The call in R is simple, we just give an argument to the *acf()* routine.

```
acf( ts, type="partial")
```

It produces the following plot for some data I created with *arima.sim()*. We have already seen that if we know that we have an autoregressive process and are looking to determine the order of the process, we produce a PACF and observe where it "cuts off": We would conclude for the time series exhibited here that we have a second order model.

```
phi.1 = .6;
phi.2 = -.6;
data.ts = arima.sim(n = 1000, list(ar = c(phi.1, phi.2)))

acf(data.ts, type="partial",
        main=paste("PACF of  Time Series Data, phi1=",phi.1,", phi.2=",phi.2) )
```



We can state another way:

> *The excess correlation at $lag = k$ not accounted for by a $(k-1)^{st}$ order model, is the partial correlation at lag=k*

Digging Into the Notation: the Conditional Correlation

> *Note: on a first reading, the casual reader (or the mathematical reader) may wish to skip the following.*

We continue to peel back on these ideas for the interested reader. To get there, let's remember an idea from basic probability. Conditional probabilities seek to get at the likelihood of an event if we know that some other, perhaps related, event has occurred.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

We can think about this in simple terms, like "your calculation of the probability that I will order a bagel increases if you see that I have just ordered coffee".

A little more deeply, consider two related random variables (X, Y) and think about their joint distributions. Start with $f(x, y)$, a joint pdf. We recall that the "conditional is the *joint* over *marginal*" (that is, the "and" over the "*given*").

> *When X is parked somewhere, say at X=x, we have a random variable $Y|X = x$, and we think of x as its parameter. The density of $Y|X = x$ will be written:*

$$f_{Y|X=x}(y)$$

OK! So when two random variables are related, and we know the first random variable has occurred with some outcome, then we don't have to chase all around for the second random variable, we just restrict our attention to the joint distribution along the outcome that we know has occurred.

For example, people's heights and weights are obviously correlated. If I know someone is 5 feet 7 inches tall (i.e. 170 cm) then I can use that information in predicting their weight. And, obviously,

the average weight for individuals who are 170 cm in height is less than the average weight for those individuals who are 200 cm (about 6 feet, 5 inches).

If you are rusty on your Calculus, you can read lightly for a few paragraphs. We recall the definition of conditional expectation (I'll write this for continuous random variables and use integrals. You can basically swap the integrals for sums in the discrete case).

$$E[Y|X = x] = \mu_{Y|X=x} = \int_{-\infty}^{\infty} y\, f_{Y|X=x}(y)\, dy$$

We'll resist the temptation to show off and recall the Total Expectation Theorem. Instead just remember that we think of variance and covariance as an expected values.

$$V[Y] = E[(Y - \mu_Y)(Y - \mu_Y)]\,, \qquad COV[X,Y] = E[(X - \mu_X)(Y - \mu_Y)]$$

Now move forward in a pretty natural way and define the *conditional variance* as "the variance of *Y* when you already know that *X=x*" (e.g. measure spread of weights for people who are 170 cm)

$$V[Y|X] = E\big[\big(Y - \mu_{Y|X=x}\big)\big(Y - \mu_{Y|X=x}\big) \mid (X = x)\big]$$

This pretty obviously depends upon (is a function of) where we park the outcome *x*.

We are almost there!

Push forward just a little more and think about a case where we have three variables in play, call them X, Y, and Z. (maybe you can think about height, weight, and age). If we know $Z = z$, we define the *conditional covariance* of *X* and *Y* given *Z=z* as well as its good friend, the partial correlation (this is just the scaled covariance)

$$COV[X,Y \mid (Z = z)] = E[(X - \mu_x)(Y - \mu_Y) \mid (Z = z)]$$

$$\rho[X,Y \mid (Z = z)] = \frac{COV[X,Y \mid Z = z]}{V[X|(Z = z)]V[Y|(Z = z)]}$$

Whew!

Bibliography

Shumway, R., & Stoffer, D. (2013). *Time Series Analysis and Its Applications.* New York: Springer.