Everybody welcome. The next topic is Multivariate Linear Models and multivariate Anova.

We'll be loading the standard packages.

Actually, it turns out that multivariate linear modelling is built into R,

it's part of the stats package.

We'll be using the fitness club example again,

and here it is.

We've already loaded it for correlations,

but I won't go through this.

Now, let's see if we want to jointly predict

the exercise measurements as a function of the physical attributes.

Turns out that doing that is very simple.

We use the LM function

and it's just only difference is that on the left-hand side we have cbind,

chins, situps, jumps,

whichever variables we want to model jointly.

That is, what cbind does is it constructs a matrix.

In this case, what it constructs is a matrix with each of these variables in columns.

Everything else is the same.

Notice that the class of the result is no longer just LM,

it's not also MLM for multivariate linear model.

Here's what the summary looks like.

Basically what it looks like is for each of the responses,

so chins, sit-ups, and jumps.

We have its own coefficient table.

This makes sense because our Beta is a matrix.

These are the corresponding rows and columns of that matrix.

Because of the way,

as we saw on the slides,

the formula works out,

you can in fact almost treat them as separate,

but where they do come together is first of all,

when we ask for the coefficients, it's a matrix.

Also for the variance covariance matrix,

it's now a matrix of every combination of response and predictor.

It's a pretty big matrix,

but that's because we have a lot of parameters here.

They're all stacked on top of each other.

We can also get a variance covariance matrix to the response.

This is, so if we were fitting individual regressions,

what we would be doing is we would be getting

these just the diagonal elements as residual variances.

However, the cover for the multivariate linear regression we also have the covariances.

Our residuals are also now matrix might put one column for each day,

each variable we're regressing and as our fitted values.

When we test hypotheses using the analysis of variance Anova,

we actually have to specify which of the multivariate Anova tests do we use.

Here it turns out to parallel very closely the canonical correlation analysis.

For example, Pillai's trace or Wilks Lambda or Hotelling test.

It basically works the same way.

We have a hypothesis matrix,

we have an error matrix and then we calculate.

There is also something called a Sphericity test, which you can perform as well.

Take a look at the help,

we won't discuss it here,

it's a repeated measures thing,

and so it will be covered in that course.

Now we can also do more complex analysis of variance.

For example, here, if we wanted to test the effect of weight given the others.

What we can do is we can reduce model without the weight,

like only waist and pulses predictors,

then call Anova with first argument as the full model.

The second argument is the reduced model,

and here's our p-value.

Would here conclude that there was not enough evidence that given waste and pulse

weight has additional predictive effect on the sports measures.

Now, unfortunately,

there are no built-in diagnostics plots for multivariate linear models.

But we can improvise,

for example, we can plot residuals.

The joint distribution of the residuals,

make sure that they are decently close to normal.

Pearson residuals, at least there is so far 3.60 don't work,

but we can make

our own Pearson residuals by getting the standard deviations and dividing through.

This is what the sweep function does, here they are.

Another approach we can take is we can decorrelate the residuals to get

some essentially constructs for

the residuals such that which contain the same information but are no longer correlated.

We do that by taking the estimated variance,

taking a Cholesky decomposition and inverting that and then

multiplying the actual residuals by this Cholesky decomposition.

Notice they're no longer correlated.

This is what they look like.

We can also make residuals versus fitted plots,

although unfortunately we have to have one for

every combination

of predictor and response,

but Douglas, we can do it and here it is.

Notice up here what we're doing is we're using the pairs function.

We are binding the PR,

which recall is the Pearson residuals

from this and the fitted values.

Then we're telling it to use the horizontal index 1-3,

vertical index 4-6,

which, horizontal index would be the PR and vertical index would be the fitted values.

Then we specify a penalty function.

Now, don't worry too much about the details,

but the basic idea is that it's emulating what R

would do by default for residuals versus fitted plot.