Take Home Exam 1

Cody Frisby

3/3/2017

1. The ABC Baseball Team

a) Unusual Players

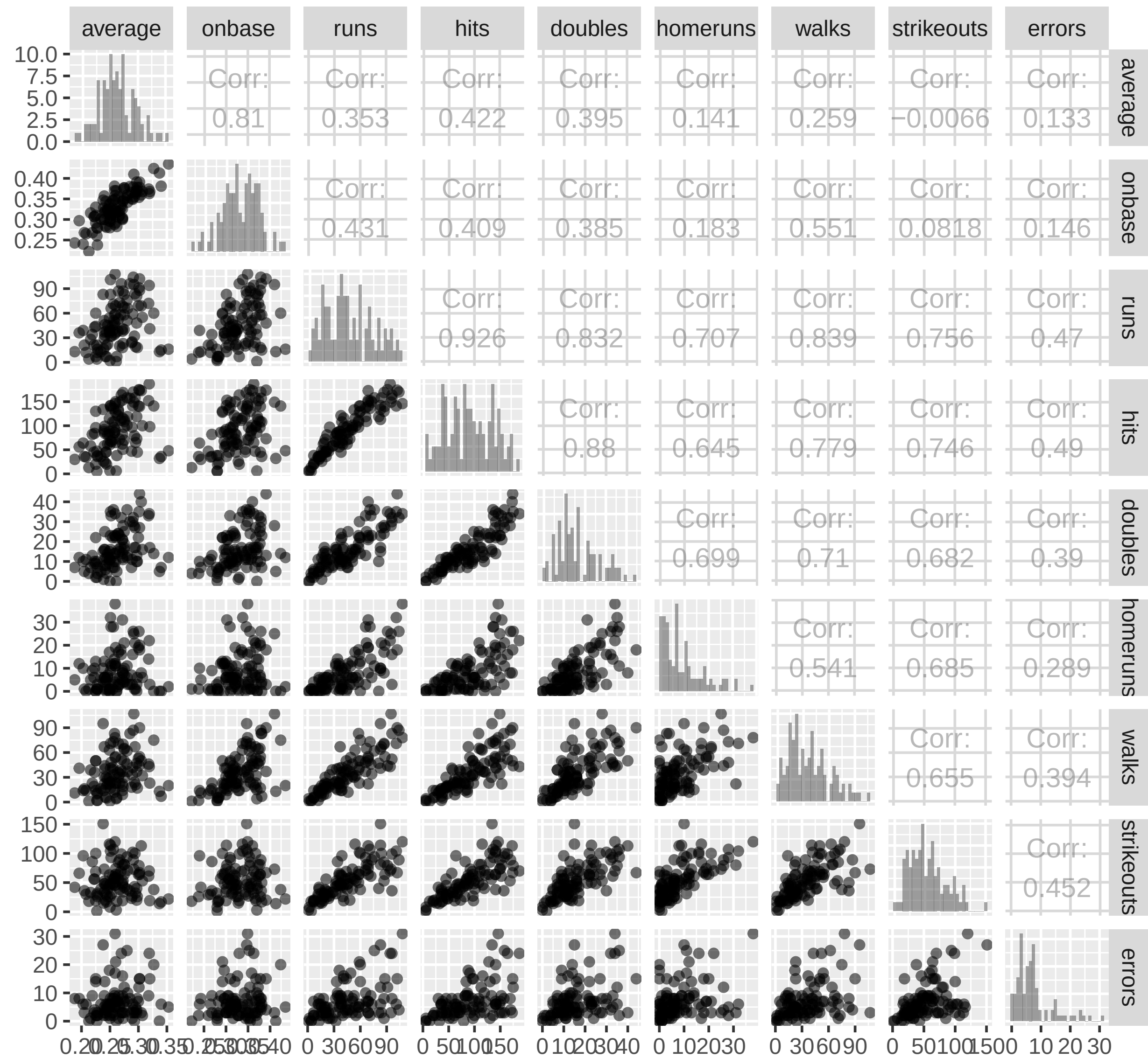
Bret Barberie is definitely a player who stands out in the two measures of batting average (0.353) and on-base percentage (0.435). Barry Bonds and Bobby Bonilla are two players that are notable for salary (6100 and 5150 respectively). Howard Johnson is notable for hitting the most home runs (38) but also for the most errors (31).

b) Notable Relationships

Number of hits and number of runs have a strong, positive relationship with $corr = 0.925697$. Other notable variable associations that have large, positive correlations (above 0.8) include on-base percentage with batting average (0.81), number of doubles with number of runs (0.832), number of doubles with number of hits (0.88), and the number of walks with number of runs (0.839).

c) Plot of the Data

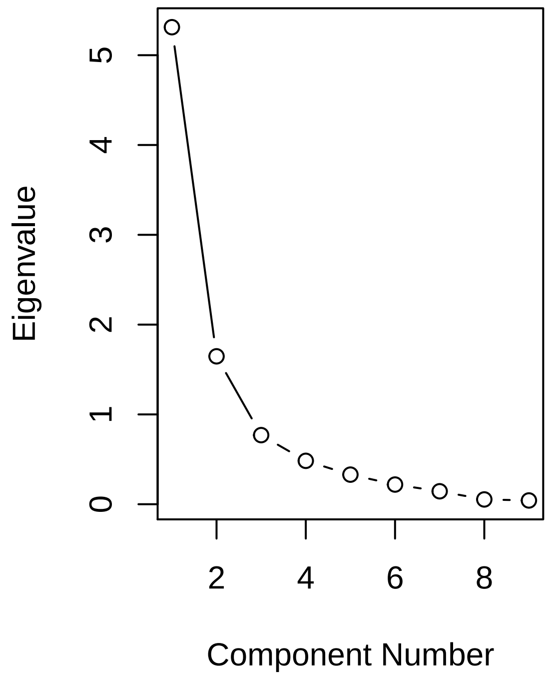
A scatter-plot matrix can succinctly, graphically summarize the data and guide us in drawing some conclusions as to what variables might be highly correlated with one-another and in aiding us with identifying some of the players mentioned in part *a*.

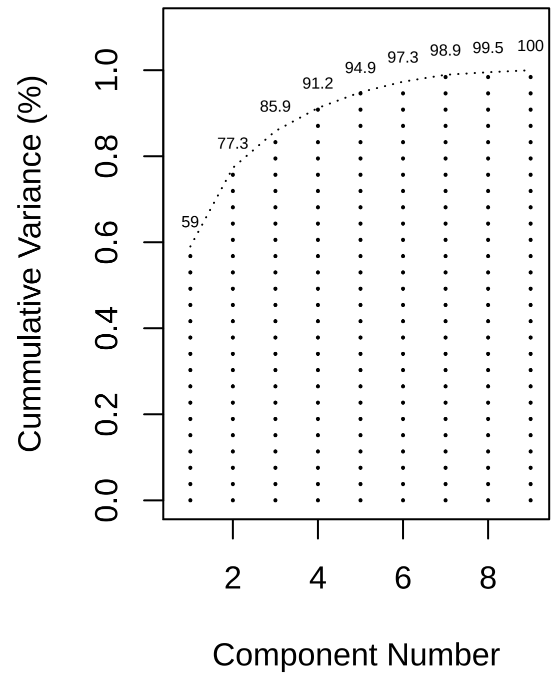


This plot can be hard to read with this many variables. It can be helpful to zoom in by looking at all the scatter-plots separately or on a smaller matrix where we remove variables that appear independent from each other.

d) Principal Components

We can describe approximately 90% of the variation in the data with 4 principal components. Here is a plot showing the relative variance of the principal components and the cumulative variance.





e) Principal Component Meanings

Those components, which describe 91.2% of the variance, are displayed here:

	PC1	PC2	PC3	PC4
average	-0.1924945	0.6445137	0.0410115	-0.4126813
onbase	-0.2275589	0.6161938	0.0052911	0.2453301
runs	-0.4138803	-0.0517425	-0.0516184	0.1026740
hits	-0.4112329	-0.0239398	0.0040676	-0.0456106
doubles	-0.3913541	-0.0265752	-0.1523248	-0.2255875
homeruns	-0.3258526	-0.2238125	-0.3500017	-0.4681886
walks	-0.3740852	0.0184842	-0.0397718	0.6751081
strikeouts	-0.3424792	-0.3614174	-0.0438554	0.0740738
errors	-0.2327740	-0.1410479	0.9200000	-0.1598028

PC_1 can be thought of as the scoring ability of a player. We have the coefficients for *runs*, *hits*, *doubles*, and *homeruns* that are larger (for the most part) than for the other principal components. PC_2 could be thought of as the batting ability of a player, or the ability of getting on base. This is indicated by the larger values for batting average and on base percentage. PC_3 overwhelmingly appears to be associated with *errors*. This component could be thought of as a player's **defense** ability. With PC_4 there appears to have one variable (like PC_3) that is strongest, *walks*. This component can be thought of as a players proneness to be walked by the opposing pitcher. There are, however, other variables that have an interesting, intuitive relationship with *walks*. As *walks* increase, *average* and *homeruns* decrease. This makes sense since a player isn't hitting the ball if they are being walked.

f) Player Level Analysis

Brent Barberie is considerably different from the other players in PC_2 respects but average for PC_1 which are the components we associated with getting on base and scoring ability respectively. Delino DeShields is a player who stands out for large value for PC_3, the "defense" component. Opposite of him, Andre Dawson, is a player who has low PC_3 and PC_4 scores. He isn't walked very often, nor does he make many *errors* which may indicate that he is an excellent defensive player.

g) Predictors of Player Salary

Fitting a linear model of the form

$$\hat{salary}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \epsilon_i$$

where the X_is, for $i = 1, 2, 3, 4$, are the first 4 principal components, we can see clearly that the first principal component is the most predictive of `salary`. Interestingly, PC_2 doesn't appear to predict *salary* very well but PC_3, the **defense** component, does.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1334.050000	99.95205	13.3468998	0.0000000
XPC1	-319.253141	43.58709	-7.3244889	0.0000000
XPC2	3.526301	78.27737	0.0450488	0.9641630
XPC3	-304.358068	114.52017	-2.6576810	0.0092316
XPC4	67.207729	144.51615	0.4650534	0.6429574

2. Mammals

1) Notable Relationships

There are notable relationships between the variables. Looking at the correlation matririx there are large values between *logbrain* and *logbody* (0.96), *logbrain* and *gestation* (0.779), as well as *logbrain* and *gestation* (0.77).

	sleeptime	lifespan	gestation	logbody	logbrain
sleeptime	1.0000000	-0.3784267	-0.5895245	-0.5512021	-0.5644951
lifespan	-0.3784267	1.0000000	0.6394415	0.6476452	0.7247805
gestation	-0.5895245	0.6394415	1.0000000	0.7711456	0.7791989
logbody	-0.5512021	0.6476452	0.7711456	1.0000000	0.9603793
logbrain	-0.5644951	0.7247805	0.7791989	0.9603793	1.0000000

2a) Test for Independence

If we can conclude that the correlation between the two sets of variables is zero then they are independent. The hypothesis for this is

$$H_0 : \rho_1 = \rho_2 = 0$$
$$H_A : \rho_i \neq 0$$

for at lease one ρ_i.

Using Bartlett's test as outlined in Everitt and Hothorn (2011, pg. 103) where the test statistic

$$\phi_0^2 = -\{n - \frac{1}{2}(q_1 + q_2 + 1)\} \sum_{i=1}^{s} log(1 - \lambda_i)$$

has a χ^2 distribution with $q_1 \times q_2$ degrees of freedom. We would reject the null hypothesis, $p < 0.00001$. There is evidence of correlation between the two sets.

2b) Significant Canonical Pairs

There are 2 cannonical dimensions, so I can run Bartlett's test twice, knocking off $(q_1 - 1)(q_2 - 1)$ degrees of freedom for the second test. The results are displayed here with p rounded to six places.

rho	Bartlett	df	pValue
0.8436570	65.70957	6	0.000000
0.3041353	4.85299	2	0.088346

We conclude that the first cannonical dimension is significant. The second dimension may be as well with a small value for p, less than 0.1. We get a similar result using Wilks Lambda test.

rho	WilksL	F	df1	df2	p
0.8436570	0.2615809	14.96521	6	94	0.000000
0.3041353	0.9075017	2.44623	2	48	0.097351

So, depending on our rejection value, I would conclude that there is at least one significant dimension, perhaps two but there is weak evidence of correlation for the second pair.

2c) Correlation

The correlation between canonical variate pairs, displayed above, are $0.843657, 0.3041353$ for dimensions 1 and 2 respectively.

2d) Canonical Variates

I define **weight** as the variate associated with *logbrain* and *logbody* and I define **character** as the variate associated with *sleeptime*, *lifespan*, and *gestation*.

Here are the raw cannonical variate coefficients for the **weight** variate.

	weight1	weight2
logbody	0.9481943	0.3176909
logbrain	0.9991655	0.0408455

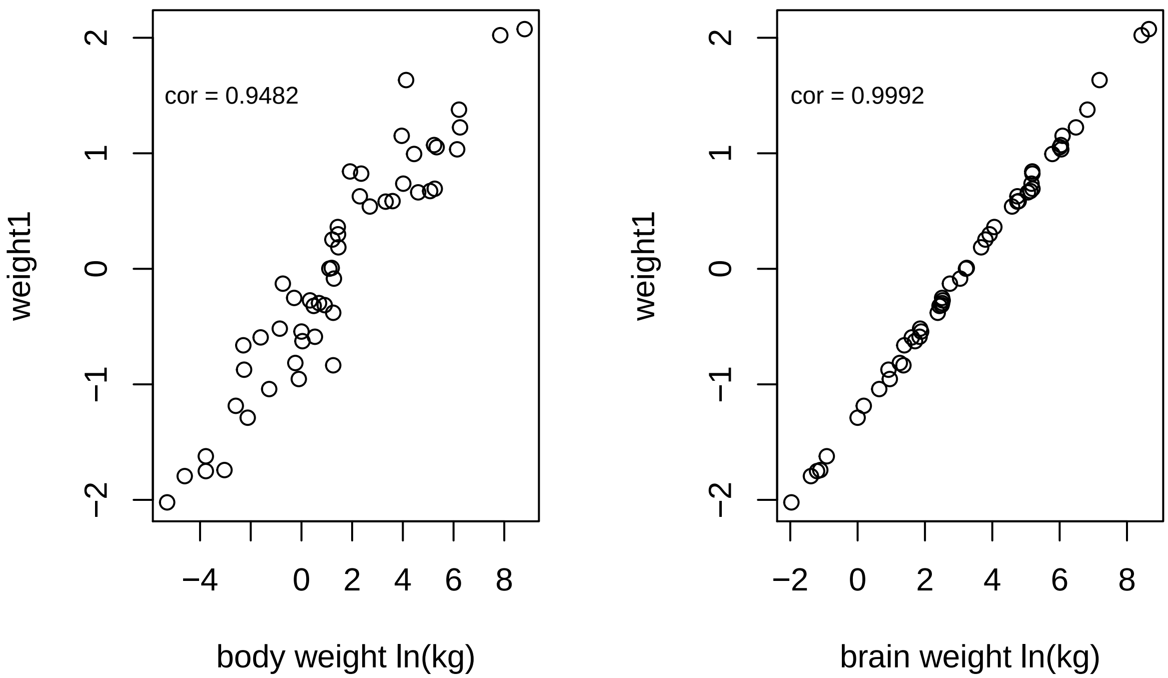
And here are those for the **character** variate.

	character1	character2
sleeptime	-0.6669704	-0.1827614

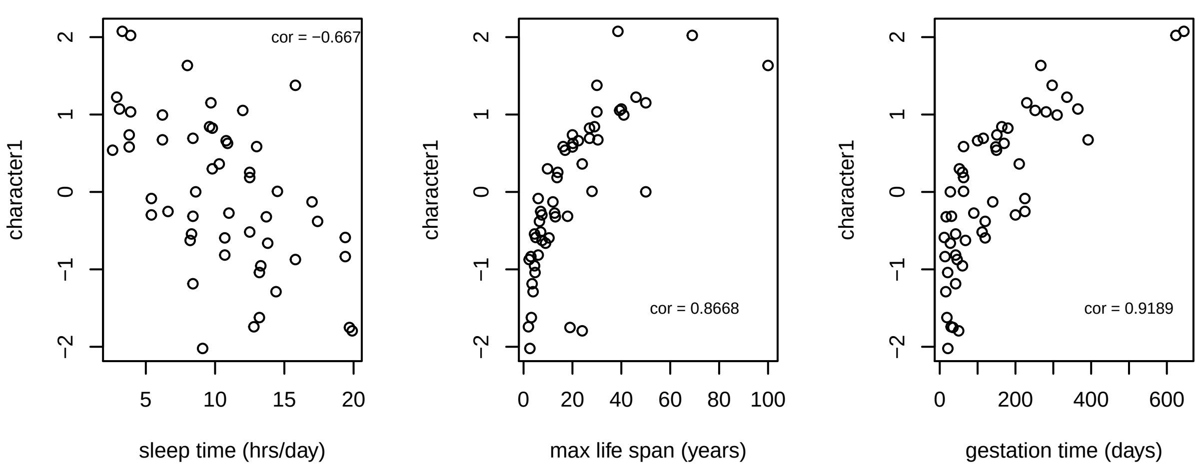
	character1	character2
lifespan	0.8667882	-0.4734240
gestation	0.9188619	0.3736384

2e) Interpret Canonical Variates

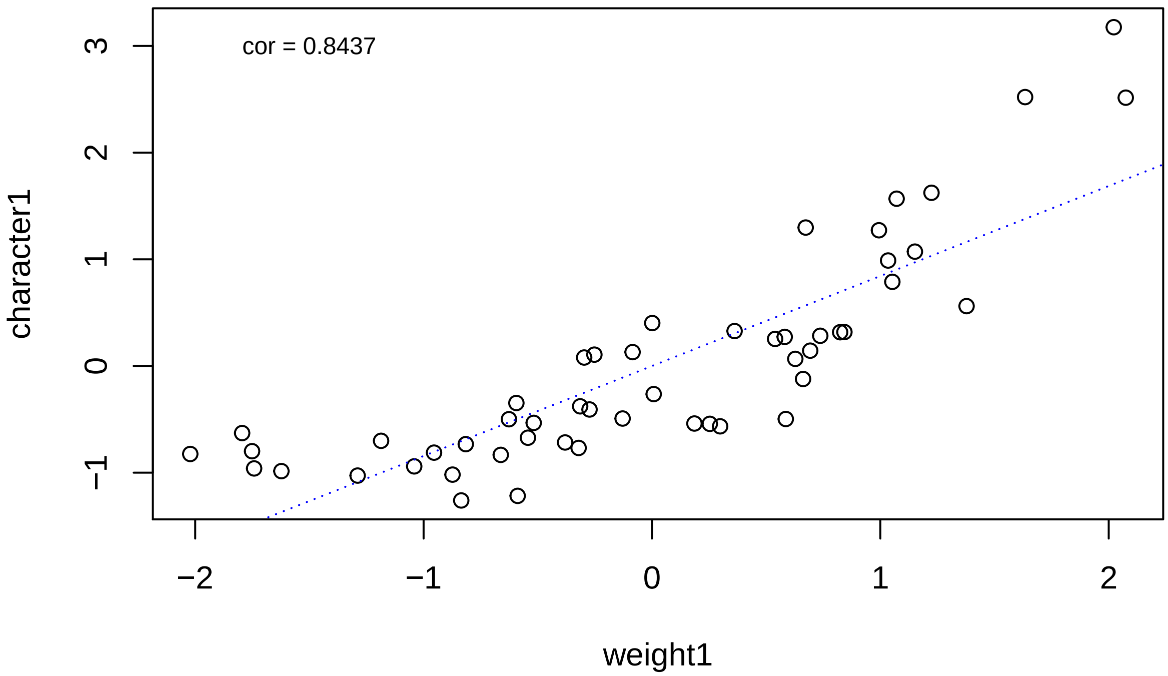
Considering **weight**, there is stong correlation between each member variable and $weight_1$. This can be thought of as the overall mass of each mammal. There is a much weaker association between the member variables and $weight_2$ (0.32 and 0.04) so it does not yield us much information about the member variables. These strong linear relationships can be illustraed with a plot.



For **character** it can be seen from the above table that there is large correlations between the member variables and $character_1$. There is a positive relationship for *lifespan* and *gestation* while there is a negative relationship with *sleeptime*. These linear relationships can also be illustrated with a scatter plot.



And, finally, a scatter plot of $weight_1$ vs. $character_1$.



NOTE: I only considered the first pair of connical variates since there is only weak evidence of correlation for the second.