```
DATA bball;
INPUT Name $     AtBats86      Hits86      HR86      Runs86      RBI86
Walks86      Years      AtBatsCar      HitsCar      HRCar      RunsCar      RBICar
WalksCar      Salary87      PutOuts86      Assists86      Errors86      Pos86 $
League  $    Team86  $    Team87 $;
CARDS;
<data goes here>
;
```

A) Why should we probably ignore the three fielding related statistics (PutOuts86, Assists86, and Errors86)?

Because they depend on the position being played (as discussed in the first question of the take home exam).

B) Create a new data set that standardizes all of the carreer statistics to be yearly averages instead of totals, and that doesn't have the fielding statistics, position, league, or teams.

```
DATA bball2;
SET bball;
DROP AtBatsCar HitsCar HRCar RunsCar RBICar WalksCar PutOuts86 Assists86 Errors86 Pos86
League Team86 Team87;
AtBatAvg = AtBatsCar/Years;
HitsAvg = HitsCar/Years;
HRAvg = HRCar/Years;
RunsAvg = RunsCar/Years;
RBIAvg = RBICar/Years;
WalksAvg = WalksCar/Years;
RUN;
```

C) One way of analyzing the data would be to simultaneously attempt to predict each of the y variables from the optimal linear combinations of the x variables. Which type of analysis is this?

This would be a multivariate multiple regression.

D) What is the best linear predictor of HR86 and RBI 86 using the standardized career numbers and years? What is the p-value for testing that all of the slope coefficients involved are 0?

```
PROC REG DATA=bball2;
MODEL HR86 RBI86 = AtBatAvg HitsAvg HRAvg RunsAvg RBIAvg WalksAvg Years;
MTEST;
RUN;
```

[NOTE: As said in class, a blank MTEST line will simultaneously test the null hypothesis that all of the coefficients for predicting the Y variables from all of the X variables are 0. If you had more Y or X variables than shown above, then you would need to list which ones you wanted to use for the test on the MTEST line.]

From the SAS output below we see that the best linear predictor of HR86 (after some rounding) is:
5.59 - 0.02 AtBatAvg + 0.04 HitsAvg + 1.26 HRAvg + 0.09 RunsAvg - 0.04 RBIAvg - 0.10 WalksAvg - 0.03 Years

The best for RBI86 is:
27.21 - 0.05 AtBatAvg - 0.20 HitsAvg - 0.31HRAvg + 0.47 RunsAvg +1.42RBIAvg - 0.36 WalksAvg - 0.43 Years

The desired p-value is < 0.0001 and we reject the null hypothesis that the set of X variables and set of Y variables are not linearly related.

Dependent Variable: HR86

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.58555	1.16476	4.80	<.0001
AtBatAvg	1	-0.02141	0.01474	-1.45	0.1478
HitsAvg	1	0.03971	0.06631	0.60	0.5499
HRAvg	1	1.26163	0.18062	6.99	<.0001
RunsAvg	1	0.08861	0.07815	1.13	0.2582
RBIAvg	1	-0.03719	0.08482	-0.44	0.6615
WalksAvg	1	-0.09636	0.03751	-2.57	0.0109
Years	1	-0.03036	0.08398	-0.36	0.7181

Dependent Variable: RBI86

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	27.20899	3.68986	7.37	<.0001
AtBatAvg	1	-0.05311	0.04670	-1.14	0.2568
HitsAvg	1	-0.19611	0.21005	-0.93	0.3516
HRAvg	1	-0.31256	0.57217	-0.55	0.5855
RunsAvg	1	0.47138	0.24759	1.90	0.0583
RBIAvg	1	1.42825	0.26870	5.32	<.0001
WalksAvg	1	-0.36021	0.11884	-3.03	0.0027
Years	1	-0.42834	0.26605	-1.61	0.1089

Multivariate Test 1

Multivariate Statistics and F Approximations

S=2 M=2 N=103.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.15655620	45.60	14	418	<.0001

E) Conduct a canonical correlation analysis for predicting the 86 batting statistics and 87 salary from the standardized career batting statistics and number of years.

This was done using PROC INSIGHT and PROC CANCOR.

```
PROC CANCORR DATA=bball2 VPREFIX=career WPREFIX=year86 ALL;
VAR  AtBatAvg HitsAvg HRAvg RunsAvg RBIAvg WalksAvg Years;
WITH AtBats86 Hits86  HR86  Runs86  RBI86  Walks86  Salary87;
RUN;
```

What is the p-value for testing whether there is any linear relationship between the x variables and the y variables?

Multivariate Statistics and F Approximations

S=7 M=-0.5 N=101

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00354293	43.28	49	1040.1	**<.0001**
Pillai's Trace	3.54520445	30.79	49	1470	<.0001
Hotelling-Lawley Trace	10.34665162	42.78	49	676.05	<.0001
Roy's Greatest Root	3.48908766	104.67	7	210	<.0001

How many of the pairs of canonical variates are significantly correlated? Report the test statistic and measure of effect size (that is, measure of how correlated).

All seven of the pairs of canonical variates are significantly correlated as can be seen from the p-values.

Test of H0: CanCorr[j]=0, j>=K

K	L. Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.003543	43.2777	49	1040.0957	**<.0001**
2	0.015905	39.3237	36	902.9789	**<.0001**
3	0.056030	35.9538	25	766.7580	**<.0001**
4	0.152118	33.7180	16	633.0331	**<.0001**
5	0.349713	30.3749	9	506.3680	**<.0001**
6	0.610826	29.2080	4	418.0000	**<.0001**
7	0.871233	31.0376	1	210.0000	**<.0001**

The estimated correlations range from 0.359 to 0.882.

Canonical Correlations

	CanCorr	Adj. CanCorr	Approx Std. Error	CanRsq
1	**0.881611**	0.869690	0.015122	0.777238
2	**0.846253**	0.837133	0.019269	0.716145
3	**0.794773**	0.784278	0.025004	0.631664
4	**0.751678**	.	0.029528	0.565020
5	**0.653817**	0.651604	0.038865	0.427476
6	**0.546713**	.	0.047594	0.298895
7	**0.358841**	.	0.059143	0.128767

Provide interpretations for the statistically significant canonical variates.

Correlations Between the VAR Variables and Their Canonical Variables

	career1	career2	career3	career4	career5	career6	career7
AtBatAvg	0.2549	-0.2662	0.5117	0.4919	0.2330	-0.5487	0.0705
HitsAvg	0.2568	-0.3139	0.6371	0.4479	0.1553	-0.4431	0.0925
HRAvg	0.8474	-0.3156	-0.0236	0.3873	0.0318	-0.1488	-0.0927
RunsAvg	0.3169	-0.1326	0.5410	0.6655	0.1433	-0.3438	0.0879
RBIAvg	0.6712	-0.4019	0.3641	0.3391	0.2714	-0.2551	-0.0414
WalksAvg	0.5794	0.3672	0.5101	0.3519	0.1856	-0.3331	0.0082
Years	0.3728	-0.0503	0.1806	-0.0873	0.2317	-0.0695	0.8716

Correlations Between the WITH Variables and Their Canonical Variables

	year861	year862	year863	year864	year865	year866	year867
AtBats86	0.1583	-0.3377	0.4223	0.5515	0.2358	-0.4932	-0.2819
Hits86	0.1405	-0.4011	0.5992	0.4809	0.0976	-0.3481	-0.3136
HR86	0.7528	-0.3888	-0.1052	0.4515	-0.0706	-0.0657	-0.2405
Runs86	0.2463	-0.1639	0.4645	0.7462	0.0788	-0.1527	-0.3322
RBI86	0.5970	-0.4869	0.2210	0.3743	0.3047	-0.1024	-0.3381
Walks86	0.5697	0.4043	0.4841	0.3226	0.2115	-0.2780	-0.2270
Salary87	0.4715	-0.1810	0.5742	0.2333	0.1603	0.0125	0.5787

The following are fairly rough interpretations (0.5 is generally used as a cut off when possible, but not always)... I wouldn't expect you to get the titles in bold unless you are a baseball person.

Power Hitter Numbers (Home Runs, RBIs, and Walks)
Career 1 = Combination of Career Averages of Home Runs, RBIs, and Walks
Year86 1 = Combination of 1986 Home Runs, RBIs, and Walks (and possibly Salary in 87)

Walks vs. Everything Else
Career 2 = Career Average in Walks contrasted with all the other Career Averages (not Years)
Year86 2 = 1986 Totals in Walks contrasted with all the other 1986 numbers (but not Salary in 87)

Getting On Base
Career 3 = Career Average in At Bats, Hits, Runs, and Walks (Hits most)
Year 86 3 = 1986 Totals in At Bats, Hits, Runs, Walks, and Salary (Hits and Salary Most)

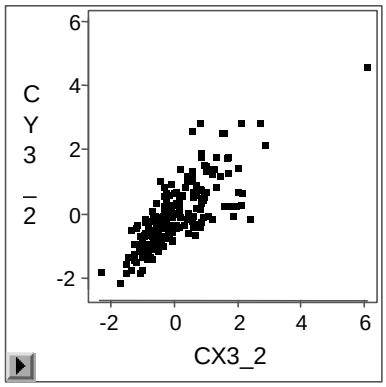
Scoring Runs
Career 4 = Combination of all the Career Averages, but especially Runs (not Years)
Year86 4 = Combination of all 1986 numbers and Salary, but especially Runs

Counting Everything But Home Runs
Career 5 = Weak average of everything but Home Runs
Year86 5 = Weak average of everything but Home Runs (Home Runs negative, Hits and Runs little weight)

????
Career 6 = At Bats and a little bit of everything else?
Year86 6 = At Bats and a little bit of everything else but Salary?

Overpaid Old Guys
Career 7 = Years in league
Year86 7 = Salary vs. everything else

Are there any points in the plots of the canonical variates that are outliers? Which players are they?



◄ Player 210 = Wade Boggs

▼Player 141 = Mike Schmidt