**CSC334/424: Assignment #2**
Due: Thursday, February 7th, 2013, by 11:59pm
Total: 65 points **(no late assignments for this assignment)**

**Problem #1 (Regression analysis - 20 points)** The Housing dataset (under the course documents for week 3) contains housing values in the suburbs of Boston. The detailed explanation concerning the input and output variables can be fetched from the UCI machine learning repository http://archive.ics.uci.edu/ml/datasets/Housing:

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centres
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of African Americans by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's (output variable)

a. Fit a linear regression model and report goodness of fit, the utility of the model, the estimated coefficients, their standard errors, and statistical significance. Use the default method for running regression analysis in SPSS and interpret your results.

**Model Summary**[b]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .861[a] | .741 | .734 | 4.7453 |

a. Predictors: (Constant), LSTAT, CHAS, B, PTRATIO, ZN, CRIM, RM, INDUS, AGE, RAD, DIS, NOX, TAX

b. Dependent Variable: MEDV

**ANOVA**[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 31637.511 | 13 | 2433.655 | 108.077 | .000[b] |
| | Residual | 11078.785 | 492 | 22.518 | | |
| | Total | 42716.295 | 505 | | | |

a. Dependent Variable: MEDV

b. Predictors: (Constant), LSTAT, CHAS, B, PTRATIO, ZN, CRIM, RM, INDUS, AGE, RAD, DIS, NOX, TAX

| Model | | Unstandardized Coefficients B | Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 36.459 | 5.103 | | 7.144 | .000 |
| | CRIM | -.108 | .033 | -.101 | -3.287 | .001 |
| | ZN | .046 | .014 | .118 | 3.382 | .001 |
| | INDUS | .021 | .061 | .015 | .334 | .738 |
| | CHAS | 2.687 | .862 | .074 | 3.118 | .002 |
| | NOX | -17.767 | 3.820 | -.224 | -4.651 | .000 |
| | RM | 3.810 | .418 | .291 | 9.116 | .000 |
| | AGE | .001 | .013 | .002 | .052 | .958 |
| | DIS | -1.476 | .199 | -.338 | -7.398 | .000 |
| | RAD | .306 | .066 | .290 | 4.613 | .000 |
| | TAX | -.012 | .004 | -.226 | -3.280 | .001 |
| | PTRATIO | -.953 | .131 | -.224 | -7.283 | .000 |
| | B | .009 | .003 | .092 | 3.467 | .001 |
| | LSTAT | -.525 | .051 | -.407 | -10.347 | .000 |

a. Dependent Variable: MEDV

$$\hat{y} = 36.45 - 0.108x_1 + 0.046x_2 + 0.021x_3 + 2.687x_4 - 17.767x_5 + 3.81x_6 + 0.001x_7 - 1.476x_8 + 0.306x_9 - 0.012x_{10} - 0.953x_{11} + 0.009x_{12} - 0.525x_{13}$$

where $y$ = MEDV $x_1$ = CRIM $x_2$ = ZN $x_3$ = INDUS $x_4$ = CHAS $x_5$ = NOX $x_6$ =RM $x_7$ =AGE $x_8$ = DIS $x_9$ =RAD $x_{10}$ = TAX $x_{11}$ =PTRATIO $x_{12}$ =B $x_{13}$=LSTA

The overall model was statistically significant and had an adjusted R^2 of .734, meaning over 73% of the variance is accounted for by the model. All the predictors except INDUS and AGE are significant.

b. Perform a feature selection on this data by using the forward selection method of the regression analysis. Analyze the output in terms of the order in which the variables are included in the regression model.

**Model Summary**[l]

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .738[a] | .544 | .543 | 6.2158 |
| 2 | .799[b] | .639 | .637 | 5.5403 |
| 3 | .824[c] | .679 | .677 | 5.2294 |
| 4 | .831[d] | .690 | .688 | 5.1386 |
| 5 | .841[e] | .708 | .705 | 4.9939 |
| 6 | .846[f] | .716 | .712 | 4.9326 |
| 7 | .850[g] | .722 | .718 | 4.8818 |
| 8 | .852[h] | .727 | .722 | 4.8474 |
| 9 | .854[i] | .729 | .724 | 4.8326 |
| 10 | .857[j] | .734 | .729 | 4.7895 |
| 11 | .861[k] | .741 | .735 | 4.7362 |

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 23243.914 | 1 | 23243.914 | 601.618 | .000[b] |
| | Residual | 19472.381 | 504 | 38.636 | | |
| | Total | 42716.295 | 505 | | | |
| 2 | Regression | 27276.986 | 2 | 13638.493 | 444.331 | .000[c] |
| | Residual | 15439.309 | 503 | 30.694 | | |
| | Total | 42716.295 | 505 | | | |
| 3 | Regression | 28988.310 | 3 | 9662.770 | 353.345 | .000[d] |
| | Residual | 13727.985 | 502 | 27.347 | | |
| | Total | 42716.295 | 505 | | | |
| 4 | Regression | 29487.388 | 4 | 7371.847 | 279.184 | .000[e] |
| | Residual | 13228.908 | 501 | 26.405 | | |
| | Total | 42716.295 | 505 | | | |
| 5 | Regression | 30246.951 | 5 | 6049.390 | 242.571 | .000[f] |
| | Residual | 12469.344 | 500 | 24.939 | | |
| | Total | 42716.295 | 505 | | | |
| 6 | Regression | 30575.223 | 6 | 5095.870 | 209.441 | .000[g] |
| | Residual | 12141.073 | 499 | 24.331 | | |
| | Total | 42716.295 | 505 | | | |
| 7 | Regression | 30848.060 | 7 | 4406.866 | 184.915 | .000[h] |
| | Residual | 11868.236 | 498 | 23.832 | | |
| | Total | 42716.295 | 505 | | | |
| 8 | Regression | 31037.996 | 8 | 3879.749 | 165.113 | .000[i] |
| | Residual | 11678.299 | 497 | 23.498 | | |
| | Total | 42716.295 | 505 | | | |
| 9 | Regression | 31132.708 | 9 | 3459.190 | 148.120 | .000[j] |
| | Residual | 11583.588 | 496 | 23.354 | | |
| | Total | 42716.295 | 505 | | | |
| 10 | Regression | 31361.312 | 10 | 3136.131 | 136.714 | .000[k] |
| | Residual | 11354.983 | 495 | 22.939 | | |
| | Total | 42716.295 | 505 | | | |
| 11 | Regression | 31634.931 | 11 | 2875.903 | 128.206 | .000[l] |
| | Residual | 11081.364 | 494 | 22.432 | | |
| | Total | 42716.295 | 505 | | | |

**Coefficients**[a]

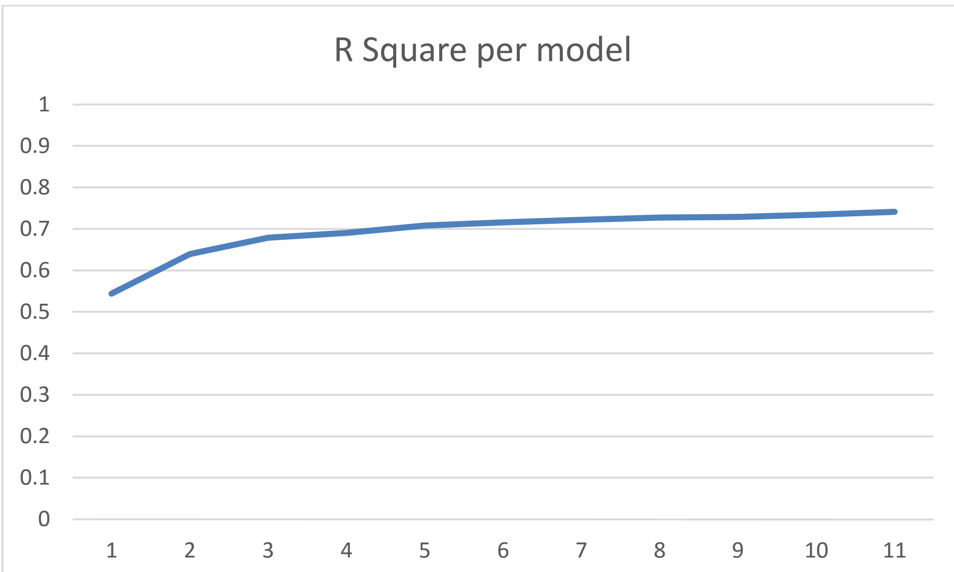
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 34.554 | .563 | | 61.415 | .000 |
| | LSTAT | -.950 | .039 | -.738 | -24.528 | .000 |
| 2 | (Constant) | -1.358 | 3.173 | | -.428 | .669 |
| | LSTAT | -.642 | .044 | -.499 | -14.689 | .000 |
| | RM | 5.095 | .444 | .389 | 11.463 | .000 |
| 3 | (Constant) | 18.567 | 3.913 | | 4.745 | .000 |
| | LSTAT | -.572 | .042 | -.444 | -13.540 | .000 |
| | RM | 4.515 | .426 | .345 | 10.603 | .000 |
| | PTRATIO | -.931 | .118 | -.219 | -7.911 | .000 |
| 4 | (Constant) | 24.471 | 4.078 | | 6.001 | .000 |
| | LSTAT | -.665 | .047 | -.517 | -14.233 | .000 |
| | RM | 4.224 | .424 | .323 | 9.966 | .000 |
| | PTRATIO | -.974 | .116 | -.229 | -8.391 | .000 |
| | DIS | -.552 | .127 | -.126 | -4.348 | .000 |
| 5 | (Constant) | 37.499 | 4.613 | | 8.129 | .000 |
| | LSTAT | -.581 | .048 | -.451 | -12.122 | .000 |
| | RM | 4.163 | .412 | .318 | 10.104 | .000 |
| | PTRATIO | -1.046 | .114 | -.246 | -9.212 | .000 |
| | DIS | -1.185 | .168 | -.271 | -7.034 | .000 |
| | NOX | -17.997 | 3.261 | -.227 | -5.519 | .000 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | (Constant) | 36.923 | 4.559 | | 8.099 | .000 |
| | LSTAT | -.570 | .047 | -.442 | -12.010 | .000 |
| | RM | 4.112 | .407 | .314 | 10.097 | .000 |
| | PTRATIO | -1.003 | .113 | -.236 | -8.895 | .000 |
| | DIS | -1.145 | .167 | -.262 | -6.865 | .000 |
| | NOX | -18.740 | 3.227 | -.236 | -5.807 | .000 |
| | CHAS | 3.244 | .883 | .090 | 3.673 | .000 |
| 7 | (Constant) | 30.412 | 4.905 | | 6.200 | .000 |
| | LSTAT | -.537 | .048 | -.417 | -11.204 | .000 |
| | RM | 4.294 | .407 | .328 | 10.561 | .000 |
| | PTRATIO | -.974 | .112 | -.229 | -8.701 | .000 |
| | DIS | -1.123 | .165 | -.257 | -6.804 | .000 |
| | NOX | -16.677 | 3.252 | -.210 | -5.129 | .000 |
| | CHAS | 3.052 | .876 | .084 | 3.484 | .001 |
| | B | .009 | .003 | .089 | 3.384 | .001 |
| 8 | (Constant) | 30.317 | 4.871 | | 6.224 | .000 |
| | LSTAT | -.543 | .048 | -.422 | -11.398 | .000 |
| | RM | 4.116 | .409 | .314 | 10.074 | .000 |
| | PTRATIO | -.882 | .116 | -.208 | -7.621 | .000 |
| | DIS | -1.383 | .188 | -.317 | -7.370 | .000 |
| | NOX | -16.687 | 3.229 | -.210 | -5.168 | .000 |
| | CHAS | 3.111 | .870 | .086 | 3.576 | .000 |
| | B | .009 | .003 | .093 | 3.563 | .000 |
| | ZN | .038 | .013 | .096 | 2.843 | .005 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | (Constant) | 29.508 | 4.873 | | 6.056 | .000 |
| | LSTAT | -.525 | .048 | -.408 | -10.858 | .000 |
| | RM | 4.150 | .408 | .317 | 10.179 | .000 |
| | PTRATIO | -.839 | .117 | -.197 | -7.147 | .000 |
| | DIS | -1.432 | .189 | -.328 | -7.591 | .000 |
| | NOX | -16.089 | 3.233 | -.203 | -4.977 | .000 |
| | CHAS | 3.030 | .868 | .084 | 3.489 | .001 |
| | B | .008 | .003 | .082 | 3.084 | .002 |
| | ZN | .042 | .013 | .107 | 3.131 | .002 |
| | CRIM | -.061 | .030 | -.057 | -2.014 | .045 |
| 10 | (Constant) | 34.712 | 5.103 | | 6.803 | .000 |
| | LSTAT | -.528 | .048 | -.410 | -11.019 | .000 |
| | RM | 3.977 | .408 | .304 | 9.754 | .000 |
| | PTRATIO | -1.015 | .129 | -.239 | -7.867 | .000 |
| | DIS | -1.429 | .187 | -.327 | -7.647 | .000 |
| | NOX | -20.314 | 3.472 | -.256 | -5.850 | .000 |
| | CHAS | 2.968 | .861 | .082 | 3.448 | .001 |
| | B | .010 | .003 | .096 | 3.591 | .000 |
| | ZN | .037 | .013 | .093 | 2.731 | .007 |
| | CRIM | -.105 | .033 | -.098 | -3.164 | .002 |
| | RAD | .129 | .041 | .122 | 3.157 | .002 |
| 11 | (Constant) | 36.341 | 5.067 | | 7.171 | .000 |
| | LSTAT | -.523 | .047 | -.406 | -11.019 | .000 |
| | RM | 3.802 | .406 | .290 | 9.356 | .000 |
| | PTRATIO | -.947 | .129 | -.223 | -7.334 | .000 |
| | DIS | -1.493 | .186 | -.342 | -8.037 | .000 |
| | NOX | -17.376 | 3.535 | -.219 | -4.915 | .000 |
| | CHAS | 2.719 | .854 | .075 | 3.183 | .002 |
| | B | .009 | .003 | .092 | 3.475 | .001 |
| | ZN | .046 | .014 | .116 | 3.390 | .001 |
| | CRIM | -.108 | .033 | -.101 | -3.307 | .001 |
| | RAD | .300 | .063 | .284 | 4.726 | .000 |
| | TAX | -.012 | .003 | -.216 | -3.493 | .001 |

$$\hat{y} = 36.341 - 0.108x_1 + 0.046x_2 + +2.719x_4 - 17.376x_5 + 3.802x_6 + -1.493x_8 + 0.3x_9 - 0.012x_{10} - 0.947x_{11} + 0.009x_{12} - 0.523x_{13}$$

The final model included only 11 of the original 13 independent variables, leaving out INDUS (proportion of non-retail business) and AGE (proportion of owner-occupied units built prior to 1940). These turn out to be the two variables which were not significant in the original model, so forward selection took care of that. All the models were significant and the R Square value improved each time (see graph below). The variables were added using forward selection, which means the best single variable model was created first, using LSTAT, the percentage of the population in lower income status. They followed in order with variables about rooms per dwelling, pupil-teacher ratio in schools, distances to jobs, air pollution, river-adjacency, ethnic minority, availability of large lots, crime, highway accessibility and property tax. While the R Squared value kept improving, the variables as they are added do not have increasingly small coefficients. In fact, PTRATIO, the third added, has a middling coefficient in the original model and RAD has a moderate coefficient originally but gets added near the end.



**Problem #2 (Canonical Correlation Analysis – 20 points):** Water, soil, and mosquito fish samples were collected at *n* = 165 sites/stations in the marshes of southern Florida. The following water variables were measured:

| | |
|---|---|
| MEHGSWB | Methyl Mercury in surface water, ng/L |
| TURB | in situ surface water turbidity |
| DOCSWD | Dissolved Organic Carbon in surface water, mg/L |
| SRPRSWFB | Soluble Reactive Phosphorus in surface water,mg/L or ug/L |
| THGFSFC | Total Mercury in mosquitofish (*Gambusia affinis*), average of 7 individuals, ug/kg |

In addition, the following soil variables were measured:

| | |
|---|---|
| THGSDFC | Total Mercury in soil, ng/g |
| TCSDFB | Total Carbon in soil, % |
| TPRSDFB | Total Phosphorus in soil, ug/g |

Perform a canonical correlation analysis, describing the relationships between the soil and water variables using the data[1] found in data_marsh_cleaned_homework#2 (both xls and spss files under the course documents for week 3).

1. Answer the following questions regarding the canonical correlations.
    a. Test the null hypothesis that the canonical correlations are all equal to zero. Give your test statistic, d.f., and p-value.

```
EFFECT .. WITHIN CELLS Regression
Multivariate Tests of Significance (S = 3, M = 1/2, N = 77 1/2)
```

| Test Name | Value | Approx. F | Hypoth. DF | Error DF | Sig. of F |
|---|---|---|---|---|---|
| Pillais | .33929 | 4.05512 | 15.00 | 477.00 | .000 |
| Hotellings | .38686 | 4.01473 | 15.00 | 467.00 | .000 |
| Wilks | .69630 | 4.05200 | 15.00 | 433.81 | .000 |
| Roys | .14868 | | | | |

Dimension Reduction Analysis

| Roots | Wilks L. | F | Hypoth. DF | Error DF | Sig. of F |
|---|---|---|---|---|---|
| 1 TO 3 | .69630 | 4.05200 | 15.00 | 433.81 | .000 |
| 2 TO 3 | .81790 | 4.17630 | 8.00 | 316.00 | .000 |
| 3 TO 3 | .92841 | 4.08707 | 3.00 | 159.00 | .008 |

[1] http://www.epa.gov/region4/sesd/reports/epa904r07001.html

The first entry of the table (in blue box) tests whether all 3 variates combined are equal to 0. We see by the last column that the p-value is significant indicating that we reject the null hypothesis that all canonical correlations are equal to 0. Note the two degrees-of-freedom values for the F test parameters. I accepted just one since we didn't discuss this in this context.

b. Test the null hypothesis that the second and third canonical correlations equal zero. Give your test statistic, d.f., and p-value.

Dimension Reduction Analysis

| Roots | Wilks L. | F | Hypoth. DF | Error DF | Sig. of F |
|---|---|---|---|---|---|
| 1 TO 3 | .69630 | 4.05200 | 15.00 | 433.81 | .000 |
| 2 TO 3 | .81790 | 4.17630 | 8.00 | 316.00 | .000 |
| 3 TO 3 | .92841 | 4.08707 | 3.00 | 159.00 | .008 |

The area in the blue box indicates the test for whether the 2nd and 3rd variates combined are significantly different from 0. We see by the last column that the p-value is significant for each of the tests indicating that we reject the null hypothesis that the 2nd and 3rd canonical correlations are equal to 0.

c. Test the null hypothesis that the third canonical correlation equals zero. Give your test statistic, d.f., and p-value.

Dimension Reduction Analysis

| Roots | Wilks L. | F | Hypoth. DF | Error DF | Sig. of F |
|---|---|---|---|---|---|
| 1 TO 3 | .69630 | 4.05200 | 15.00 | 433.81 | .000 |
| 2 TO 3 | .81790 | 4.17630 | 8.00 | 316.00 | .000 |
| 3 TO 3 | .92841 | 4.08707 | 3.00 | 159.00 | .008 |

The line in the blue box shows the test results for the final variate by itself.

d. Present the three canonical correlations

Eigenvalues and Canonical Correlations

| Root No. | Eigenvalue | Pct. | Cum. Pct. | Canon Cor. | Sq. Cor |
|---|---|---|---|---|---|
| 1 | .17464 | 45.14311 | 45.14311 | .38558 | .14868 |
| 2 | .13510 | 34.92338 | 80.06649 | .34500 | .11902 |
| 3 | .07711 | 19.93351 | 100.00000 | .26757 | .07159 |

e. What can you conclude from the above analyses?

All three canonical correlations are statistically significant and they are all larger than about 0.2. Therefore they may all be useful.

2. Answer the following questions regarding the canonical variates.
   a. Give the formulae for the significant canonical variates for the soil and water variables.

EFFECT .. WITHIN CELLS Regression (Cont.)
Univariate F-tests with (5,159) D. F.

| Variable | Sq. Mul. R | Adj. R-sq. | Hypoth. MS | Error MS | F | Sig. of F |
|---|---|---|---|---|---|---|
| THGSDFC | .10864 | .08060 | 14869.32167 | 3836.62128 | 3.87563 | .002 |
| TCSDFB | .13107 | .10375 | 651.81285 | 135.88601 | 4.79676 | .000 |
| TPRSDFB | .11193 | .08400 | 128509.30137 | 32062.97794 | 4.00803 | .002 |

Raw canonical coefficients for DEPENDENT variables
Function No.

| Variable | 1 | 2 | 3 |
|---|---|---|---|
| THGSDFC | -.01142 | -.01017 | .01411 |
| TCSDFB | .07756 | -.03772 | -.07279 |
| TPRSDFB | .00297 | .00227 | .00422 |

Standardized canonical coefficients for DEPENDENT variables
Function No.

| Variable | 1 | 2 | 3 |
|---|---|---|---|
| THGSDFC | -.73743 | -.65693 | .91123 |
| TCSDFB | .95497 | -.46446 | -.89625 |
| TPRSDFB | .55554 | .42444 | .79002 |

Raw canonical coefficients for COVARIATES
Function No.

| COVARIATE | 1 | 2 | 3 |
|---|---|---|---|
| MEHGSWB | -.72057 | -.61331 | -.44282 |
| TURB | -.01490 | .00395 | -.04659 |
| DOCSWD | .12290 | -.04565 | .03831 |
| SRPRSWFB | 15.97272 | 77.86417 | 98.95910 |
| THGFSFC | -.00412 | -.00985 | .00949 |

Standardized canonical coefficients for COVARIATES
CAN. VAR.

| COVARIATE | 1 | 2 | 3 |
|---|---|---|---|
| MEHGSWB | -.32611 | -.27757 | -.20041 |
| TURB | -.16111 | .04268 | -.50364 |
| DOCSWD | 1.05314 | -.39118 | .32826 |
| SRPRSWFB | .10646 | .51898 | .65958 |
| THGFSFC | -.26750 | -.63876 | .61571 |

Formulas:

$$SoilVariate1 = -.011 * THGSDFC + .078 * TCSDFB + .003 * TPRSDFB$$

$$SoilVariate2 = -.010 * THGSDFC - .038 * TCSDFB + .002 * TPRSDFB$$

$$SoilVariate3 = .014 * THGSDFC - .073 * TCSDFB + .004 * TPRSDFB$$

$$WaterVariate1 = -.721 * MEHGSWB - .015 * TURB + 0123 * DOCSWD + 15.97 * SRPRSWFB - .004 * THGFSFC$$

$$WaterVariate2 = -.613 * MEHGSWB + .004 * TURB - .046 * DOCSWD + 77.86 * SRPRSWFB - .010 * THGFSFC$$

$$WaterVariate3 = -.443 * MEHGSWB - .047 * TURB + .038 * DOCSWD + 98.96 * SRPRSWFB + .009 * THGFSFC$$

b. Give the correlations between the significant canonical variates for soils and the soil variables, and the correlations between the significant canonical variates for water and the water variables. (The water variables are called COVARIATES in this case by SPSS)

Correlations between DEPENDENT and canonical variables
Function No.

| Variable | 1 | 2 | 3 |
|---|---|---|---|
| THGSDFC | .00951 | -.88365 | .46806 |
| TCSDFB | .63909 | -.76826 | -.03666 |
| TPRSDFB | .71407 | .14767 | .68433 |

```
Correlations between COVARIATES and canonical variables
          CAN. VAR.

Covariate                  1                2                3

MEHGSWB               .21383          -.54424          -.05581
TURB                  .12070          -.03436          -.49853
DOCSWD                .89202          -.39006          -.02465
SRPRSWFB              .17194           .58138           .63984
THGFSFC              -.49143          -.62010           .52590
```

c. What can you conclude from the above analyses?

   For the top conical correlation, the Water variables that contribute most to the Water Cononical Variate are Docswd and Thgfsfc. The Soil variables that contribute most to the Soil Cononical variate are Tcsdfb and Tprsdfb. The correlation between the variates is not very high, but we may want to look into this relationship which suggests that the carbon and phosphorous content of the soil are correlated to the carbon in the water and negatively correlated to the mercury in the water.

   In the second variate pair, we see a potential relationship with mercury in water and mercury in fish together being negatively correlated with mercury in the soil and carbon in the soil.

**Problem 3 (Principal Component Analysis - 20 points):** The data given in the file 'problem3.txt'[2] (under course documents for week 3) is the percentage employed in different industries in Europe countries during 1979. Techniques such as Principal Component Analysis (PCA) can be used to examine which countries have similar employment patterns. There are 26 countries in the file and 10 variables as follows:

Variable Names:

1. Country: Name of country
2. Agr: Percentage employed in agriculture
3. Min: Percentage employed in mining
4. Man: Percentage employed in manufacturing
5. PS: Percentage employed in power supply industries
6. Con: Percentage employed in construction
7. SI: Percentage employed in service industries
8. Fin: Percentage employed in finance
9. SPS: Percentage employed in social and personal services

[2] http://lib.stat.cmu.edu/DASL/Datafiles/EuropeanJobs.html

10. TC: Percentage employed in transport and communications.

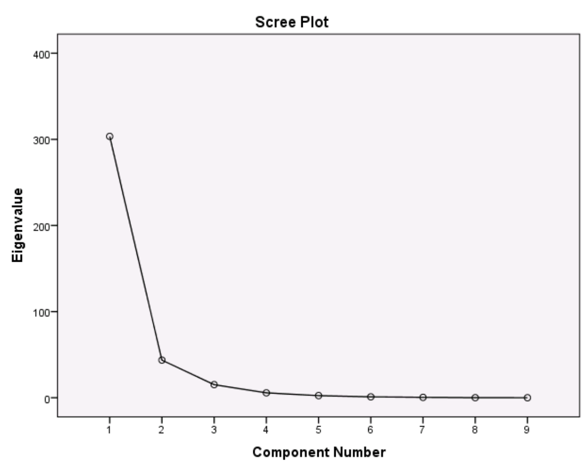
Perform a principal component analysis using the covariance matrix:

a. How many principal components are required to explain 90% of the total variation for this data?

**Total Variance Explained**

| | Component | Initial Eigenvalues[a] | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|
| | | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| Raw | 1 | 303.458 | 81.578 | 81.578 | 303.458 | 81.578 | 81.578 |
| | 2 | 43.702 | 11.748 | 93.327 | 43.702 | 11.748 | 93.327 |
| | 3 | 15.207 | 4.088 | 97.415 | | | |
| | 4 | 5.639 | 1.516 | 98.931 | | | |
| | 5 | 2.443 | .657 | 99.588 | | | |
| | 6 | 1.046 | .281 | 99.869 | | | |
| | 7 | .421 | .113 | 99.982 | | | |
| | 8 | .065 | .017 | 99.999 | | | |
| | 9 | .002 | .001 | 100.000 | | | |

You would need 2 principal components to explain 90% of the total variation




b. For the number of components in part a, give the formula for each component and a brief interpretation.

**Component Score Coefficient Matrix**[a]

| | Component | |
|---|---|---|
| | 1 | 2 |
| Agr | -.796 | -.016 |
| Min | .000 | .014 |
| Man | .109 | .817 |
| PS | .000 | .001 |
| Con | .005 | .017 |
| Sl | .050 | -.162 |
| Fin | .005 | -.055 |
| SPS | .117 | -.586 |
| TC | .004 | -.002 |

Extraction Method: Principal Component Analysis.
Component Scores.

a. Coefficients are standardized.

Equations:

$$Comp.1 = -.796Agr + .000Min + .109Man + .000PS + .005Con - .050Sl - .005Fin - .117SPS + .004TC$$

$$Comp.2 = -.016Agr - .014Min - .817Man + .001PS - .017Con - .162Sl + .055Fin + .586SPS - .002TC$$

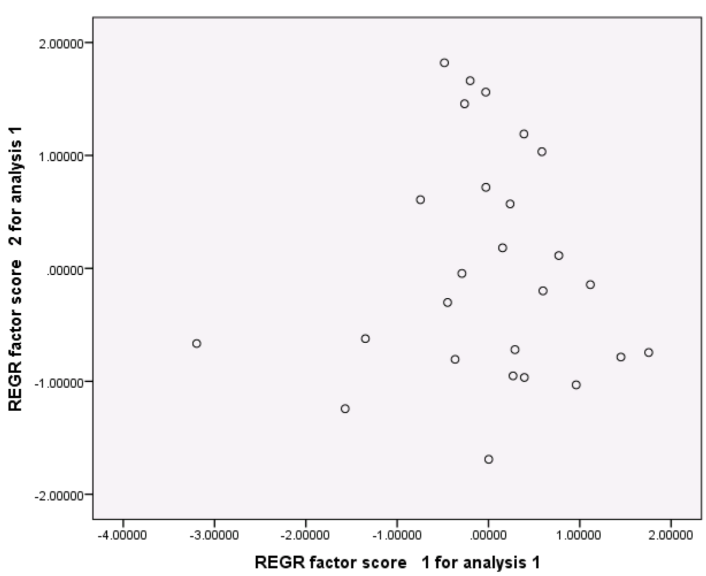
The first component is dominated by a negative component in agricultural employment. The second component corresponds highly to manufacturing employment but negatively to social and personal services jobs.

c) What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give the principal component scores (again only for the number of components specified in part a).

| Country | FAC1_1 | FAC2_1 |
|---|---|---|
| Belgium | 1.00555 | -.74519 |
| Denmark | .65997 | -1.76407 |
| France | .52403 | -.32799 |
| W. Germany | .82626 | .76353 |
| Ireland | -.25592 | -.92752 |
| Italy | .23115 | -.05883 |
| Luxembourg | .69401 | .35282 |
| Netherlands | .79796 | -1.47088 |
| United Kingdom | 1.07512 | -.50400 |
| Austria | .37149 | .50776 |
| Finland | .39248 | -.60150 |
| Greece | -1.45964 | -.27299 |
| Norway | .62985 | -1.34006 |
| Portugal | -.53983 | -.01296 |
| Spain | -.33151 | .93162 |
| Sweden | .87899 | -1.28984 |
| Switzerland | .72812 | 1.47929 |
| Turkey | -2.99170 | -1.30722 |
| Bulgaria | -.23862 | 1.01454 |
| Czechoslovakia | .18634 | 1.39692 |
| E. Germany | .99974 | 1.62347 |
| Hungary | -.18001 | .75437 |
| Poland | -.76439 | .44546 |
| Rumania | -.97654 | 1.38037 |
| USSR | -.26332 | -.13190 |
| Yugoslavia | -1.99957 | .10479 |

These come from the data table. Under *scores* in the options for dimension reduction, select the option to save scores, and these will be added as columns in the data window.

c. Include and interpret the scatter plot of the data using the first two principal components.



There is little correlation between the first two principle components. This is the idea behind a principle components analysis. You can see the core vs outer edge of the data, but tighter clusters are not apparent.

**Problem 4 (overview – 5 points):** Briefly describe the similarities and differences between:

a. Linear regression and canonical correlation

   A linear regression has only 1 dependent variable and multiple independent variables while a canonical correlation has multiple dependent variables as well as multiple independent variables. Regression is a special case of canonical correlation when there is only one dependent variable.

b. Canonical correlation and principal component analysis

   Both help to find relationships between sets of variables, but with PCA it is within a single group and with CCA it is across two groups.

   PCA can be used in Canonical correlation on the variables in the dataset to remove the correlation between the variables before running the CCA.

   PCA is mainly concerned with reducing the number of features in a dataset