CSC 424- Assignment 2

Sarah Cummings

Problem 1- Boston Housing Data

a. Fit linear regression with default SPSS method.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.861[a]	.741	.734	4.74530

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

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
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

Overall, this model seems to work pretty well.

Goodness of fit: This model has r squared of 0.741, which means that the model accounts for 0.741 % of the variability in the data.

Utility of the model: This model has a F stat of 108.077, which is statistically significant with p=0.00. Thus we can conclude the model is useful in predicting median house value.

Estimated model/ coefficients: MEDV = 36.459 -0.108 CRIM + 0.046 ZN +0.021 INDUS + 2.687 CHAS -17.767NOX +3.810 RM +0.001 AGE -1.476 DIS +0.306 RAD -0.012 TAX -0.0953 PTRATIO + 0.009B -0.525LSTAT.

Standard error and significance: As seen in the coefficient table above,
the constant has std error 5.103 and t=7.144, which is significant.
CRIM has std error 0.033 and t=-3.287.144, which is significant.
ZN has std error 0.014 and t=3.382, which is significant.
INDUS has std error 0.061and t=0.334, which is not significant.
CHAS has std error 0.862 and t=3.118, which is significant.
NOX has std error 3.820 and t= -4.651, which is significant.
RM has std error 0.418 and t=9.116, which is significant.
AGE has std error 0.013 and t=0.052, which is **not** significant.
DIS has std error 0.199 and t= -7.358, which is significant.

TAX has std error 0.004 and t=-3.280, which is significant.
PTRATIO has std error 0.131 and t= -7.283, which is significant.
B has std error 0.003 and t= 3.467, which is significant.
LSTAT has std error 0.051 and t= -10.347, which is significant.

b. Fit a regression by forward selection.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	LSTAT	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	RM	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
3	PTRATIO	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
4	DIS	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
5	NOX	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
6	CHAS	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
7	B	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
8	ZN	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
9	CRIM	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
10	RAD	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
11	TAX	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: MEDV

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
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

a. Dependent Variable: MEDV

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
11	Regression	31634.931	11	2875.903	128.206	.000'
	Residual	11081.364	494	22.432		
	Total	42716.295	505			

*Note, I have reduced the ANOVA and coefficient tables above so only the data on the final model is displayed

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.738[a]	.544	.543	6.21576
2	.799[b]	.639	.637	5.54026
3	.824[c]	.679	.677	5.22940
4	.831[d]	.690	.688	5.13858
5	.841[e]	.708	.705	4.99387
6	.846[f]	.716	.712	4.93263
7	.850[g]	.722	.718	4.88178
8	.852[h]	.727	.722	4.84743
9	.854[i]	.729	.724	4.83260
10	.857[j]	.734	.729	4.78951
11	.861[k]	.741	.735	4.73623

a. Predictors: (Constant), LSTAT

b. Predictors: (Constant), LSTAT, RM

c. Predictors: (Constant), LSTAT, RM, PTRATIO

d. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS

e. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX

f. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX, CHAS

g. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B

h. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN

i. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN, CRIM

j. Predictors: (Constant), LSTAT, RM, PTRATIO, DIS, NOX, CHAS, B, ZN, CRIM, RAD

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

Our final model chosen via forward selection has 11 of the 13 independent variables. See the coefficient table (above) for the final coefficients and variables of the model.

Analyze output: With forward selection, the most useful/ significant variable enters at each step until no further variable brings significant utility to the model. LSTAT enters the model first, being the most significant variable, and it is followed by RM, PTRATIO, DIS, NOX, CHAS, B, ZN, CRIM, RAD, and TAX. As seen in the model summary table, the adjusted r-squared increases with each additional variable added to the model. As we near the 11th model, the increase of adjusted r-squared tapers off. Our final model has adjusted r-squared of 0.735, which means our model accounts fro 73.5% of the variability in our data.

Problem 2- Canonical Correlation Analysis

Syntax input:

```
MANOVA  THGSDFC TCSDFB TPRSDFB  WITH MEHGSWB TURB DOCSWD SRPRSWFB THGFSFC
/discrim all alpha(1)
/print=sig(eigen dim).
```

1. Questions about canonical correlations:
   a) Test null hypothesis that canonical correlations are equal to zero.

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**

Because of the way the eigenvectors are computed, we can test the null hypothesis that canonical correlations are equal to zero by looking at the significance of the first F statistic in the Dimension Reduction Analysis. As seen in the first line above, the F=4.052 which corresponds to p=0.000 with df=15. Thus, we reject the null hypothesis that canonical correlations are equal to zero.

b) Test the null hypothesis that the second and third canonical correlations equal to zero.

As seen in the second line above, the F= 4.17630 which corresponds to p= 0.00 with df=8. Thus, we reject the null hypothesis that the second and third canonical correlations are equal to zero.

c) Test the null hypothesis that the third canonical correlation equals zero.

As seen in the third line above, the F=4.08707 which corresponds to p=00.008 with df=3. Thus, we reject the null hypothesis that the third canonical correlation equals zero.

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

The three canonical correlations are 0.38558, 0.34500 and 0.26757 as seen in the output above.

e) Conclusions about the Canonical correlations:
From part a, we can conclude that our canonical correlations are significantly different from zero. This means there is evidence of correlation between our water and soil variables. However, looking at the canonical correlation values (part d), we see that each canonical correlation is around the 0.3 range, and thus we do not have a particularly strong correlation.

2. Questions about the canonical variates.
   a) Formula for the significant canonical variates:

Raw canonical coefficients for DEPENDENT variables
Function No.

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

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**

The formulas for the soil canonical variates can be derived from the output above.
C1= -0.01142 THGSDFC + 0.07756 TCSDFB + 0.00297 TPRSDFB
C2= -0.01017 THGSDFC -0.03772 TCSDFB +0.00227 TPRSDFB
C3 = 0.01411 THGSDFC -0.07279 TCSDFB + 0.00422 TPRSDFB

and the other three formulas for the water variates can be formed similarly from the bolded values above, with each formula consisting of four coefficients and four variables.

b) 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:

```
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
```

For the soil canonical varaites (above), C1 has a particularly strong positive correlation with both TCSDFB (0.63909) and TPRSDFB (0.71407) . C2 has a strong negative correlation with THGSDFC (-0.88365) and TCSDFB (-0.76826). C3 has a moderate positive relationship with THGSDFC (0.46806) and a strong positive relationship with TPRSDFB (0.68433).

For the water canonical variates (above), also see the bolded correlation values for noteworthy correlations in each variate.

c) Conclusions about the analysis:

From our canonical correlation analysis, we can conclude that there is a significant relationship between the water and soil variables for this data. We were able to form t three soil canonical variates based on original three soil variables, and three water canonical variates based on the original four water variables. These variates were all significant, and they each help us understand the relationships between the variables in our data.

Problem 3- Principal Component Analysis

a) Number of principal components needed to explain 90% of the total variation for this data : 5. Looking at the Total Variance Explained table (below, left), we see that with 5 components, 91.711% of the variance is explained.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.487	38.746	38.746	3.487	38.746	38.746	2.413	26.807	26.807
2	2.130	23.669	62.415	2.130	23.669	62.415	1.732	19.248	46.055
3	1.099	12.211	74.625	1.099	12.211	74.625	1.537	17.079	63.133
4	.994	11.050	85.675	.994	11.050	85.675	1.459	16.215	79.348
5	.543	6.036	91.711	.543	6.036	91.711	1.113	12.363	91.711
6	.383	4.260	95.971						
7	.226	2.508	98.480						
8	.137	1.520	99.999						
9	4.563E-5	.001	100.000						

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component				
	1	2	3	4	5
Agr	-.978	.078	-.051	.029	.157
Min	-.002	.902	.211	.064	-.121
Man	.649	.518	.158	-.345	-.284
PS	.478	.381	.588	.392	.218
Con	.607	.075	-.161	-.666	.348
SI	.708	-.511	.121	-.050	-.209
Fin	.139	-.662	.616	-.051	.206
SPS	.723	-.323	-.327	.411	-.162
TC	.685	.296	-.393	.314	.378

Extraction Method: Principal Component Analysis.

a. 5 components extracted.

b) Formula for each component from a:

C1= -0.978Agr -.002Min +.649 Man +.478PS +.607Con +.708SI + .139Fin + .723SPS +.685TC
And C2-C5 are formed similarly from the component matrix seen above (right):

Interpretation: C1 includes a strong negative Agr coefficient, and strong positive coefficients for Man, PS, Con, SI, SPS and TC. C2's most note worthy coefficients are positive coefficients for Min and Man, and negative coefficients for SI and Fin. See the component matrix above (right) for a breakdown of the other components and their coefficients.

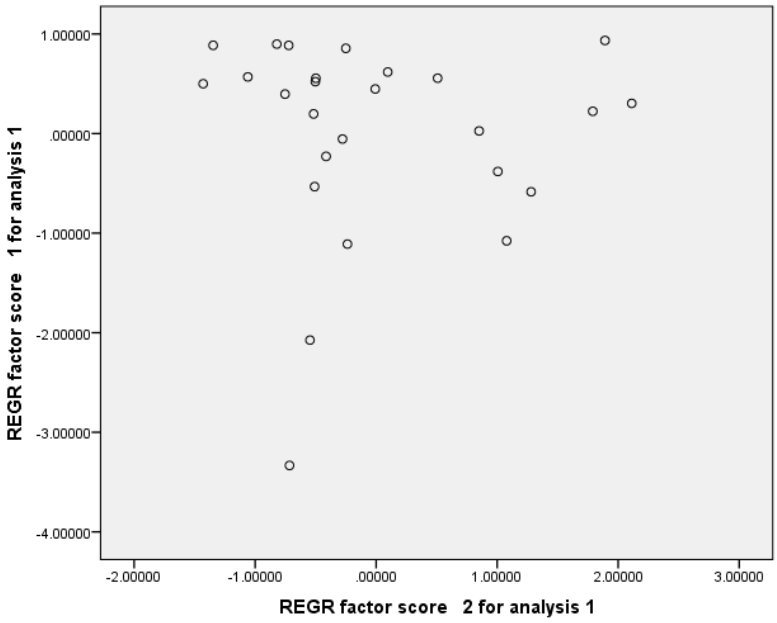
c) Which countries with highest and lowest values for each principal component:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
REGR factor score 1 for analysis 2	26	-3.33314	.93512	.0000000	1.00000000
REGR factor score 2 for analysis 2	26	-1.42956	2.11194	.0000000	1.00000000
REGR factor score 3 for analysis 2	26	-2.26645	2.91125	.0000000	1.00000000
REGR factor score 4 for analysis 2	26	-2.45457	1.27959	.0000000	1.00000000
REGR factor score 5 for analysis 2	26	-1.56771	1.85707	.0000000	1.00000000
Valid N (listwise)	26				

For component 1, minimum is -3.333 (Turkey) and maximum is 0.93512 (E. Germany).
For component 2, minimum is -1.429 (Denmark) and max is 2.111 (Hungary).
For component 3, minimum is -2.266 (USSR) and max is 2.911 (Yugoslavia).
For component 4, minimum is -2.454 (Spain) and max is 1.279 (Hungary).
For component 5, minimum is -1.567 (Turkey) and max is 1.857 (Spain).

d) Scatterplot of first and second components:



Interpretation: I'm not really sure how to interpret the scatterplot of two PCA components. It looks like there is not a definitive pattern between the two components; there is no real correlation between a country's score for component 1 versus component 2. I suppose this is a good thing, and these components are supposed to be independent and orthogonal.

Problem 4- Overview/ Similarities and differences:

a) Linear regression and Canonical Correlation:

Linear regression is used to a predict quantitative dependent variable with quantitative and categorical independent variables. Selection methods such has forward selection, backward elimination, and stepwise selection allow us to reduce dimensionality and use only the independent variables that are most useful to us.

Rather than focusing on the prediction of one dependent variable, CCA examines the relationship of two sets of variables such that there are multiple independent variables and multiple dependent variables. In the analysis, we create linear composites of the dependent variables and independent variables and form new variates that best represents the Xs and new variates that best represents the Ys.

Linear regression and canonical correlation can both be used for prediction and both are methods of supervised learning. Multiple linear regression can actually also be considered a type of canonical analysis.

b) CCA and PCA:

Like CCA, PCA is another supervised learning method that helps us discover patterns in our variables and data. After discovering these patterns, PCA allows us to compress the data and reduce the dimensionality by forming new orthogonal vectors, called components, such that the the input data are a linear combination of the principal components. Another similarity between the two is that both methods come up with new functions formed from linear regressions of the original variables. These are the variates (formed from original variables in CCA) and the components (created from normalized input data in PCA).