**CSC424/334: Assignment #2**

Due: June 28, 2017 by 11:59pm

Total: 35 points

**Problem #1** **(Regression analysis - 15 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 blacks 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 Summaryb** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .861a | .741 | .734 | 4.7453 |
| a. Predictors: (Constant), lstat, chas, black, ptratio, zn, crim, rm, indus, age, rad, dis, nox, tax | | | | |
| b. Dependent Variable: medv | | | | |

The goodness of fit is given by the F-Value which is **108.077** and the p-value associated with the F-Value is < 0.05 meaning the model is a good model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 31637.511 | 13 | 2433.655 | 108.077 | .000b |
| Residual | 11078.785 | 492 | 22.518 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| a. Dependent Variable: medv | | | | | | |
| b. Predictors: (Constant), lstat, chas, black, ptratio, zn, crim, rm, indus, age, rad, dis, nox, tax | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
| B | Std. Error | Beta | Lower Bound | Upper Bound |
| 1 | (Constant) | 36.459 | 5.103 |  | 7.144 | .000 | 26.432 | 46.487 |
| crim | -.108 | .033 | -.101 | -3.287 | .001 | -.173 | -.043 |
| zn | .046 | .014 | .118 | 3.382 | .001 | .019 | .073 |
| indus | .021 | .061 | .015 | .334 | .738 | -.100 | .141 |
| chas | 2.687 | .862 | .074 | 3.118 | .002 | .994 | 4.380 |
| nox | -17.767 | 3.820 | -.224 | -4.651 | .000 | -25.272 | -10.262 |
| rm | 3.810 | .418 | .291 | 9.116 | .000 | 2.989 | 4.631 |
| age | .001 | .013 | .002 | .052 | .958 | -.025 | .027 |
| dis | -1.476 | .199 | -.338 | -7.398 | .000 | -1.867 | -1.084 |
| rad | .306 | .066 | .290 | 4.613 | .000 | .176 | .436 |
| tax | -.012 | .004 | -.226 | -3.280 | .001 | -.020 | -.005 |
| ptratio | -.953 | .131 | -.224 | -7.283 | .000 | -1.210 | -.696 |
| black | .009 | .003 | .092 | 3.467 | .001 | .004 | .015 |
| lstat | -.525 | .051 | -.407 | -10.347 | .000 | -.624 | -.425 |
| a. Dependent Variable: medv | | | | | | | | |

The Regression Equation is as follows:

Medv = 36.459 - 0.108\*crim + 0.046\*zn + 0.021\*indus + 2.687\*chas – 17.76\*nox + 3.810\*rm + 0.001\*age – 1.476\*dis + 0.306\*rad – 0.012\*tax – 0.953\*ptratio + 0.009\*black – 0.525\*lstat

As we can see except for the “**indus**” and “**age**” variables all other variables are significant for the model.

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 Summaryl** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .738a | .544 | .543 | 6.2158 |
| 2 | .799b | .639 | .637 | 5.5403 |
| 3 | .824c | .679 | .677 | 5.2294 |
| 4 | .831d | .690 | .688 | 5.1386 |
| 5 | .841e | .708 | .705 | 4.9939 |
| 6 | .846f | .716 | .712 | 4.9326 |
| 7 | .850g | .722 | .718 | 4.8818 |
| 8 | .852h | .727 | .722 | 4.8474 |
| 9 | .854i | .729 | .724 | 4.8326 |
| 10 | .857j | .734 | .729 | 4.7895 |
| 11 | .861k | .741 | .735 | 4.7362 |
| 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, black | | | | |
| h. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn | | | | |
| i. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim | | | | |
| j. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad | | | | |
| k. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad, tax | | | | |
| l. Dependent Variable: medv | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 23243.914 | 1 | 23243.914 | 601.618 | .000b |
| Residual | 19472.381 | 504 | 38.636 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 2 | Regression | 27276.986 | 2 | 13638.493 | 444.331 | .000c |
| Residual | 15439.309 | 503 | 30.694 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 3 | Regression | 28988.310 | 3 | 9662.770 | 353.345 | .000d |
| Residual | 13727.985 | 502 | 27.347 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 4 | Regression | 29487.388 | 4 | 7371.847 | 279.184 | .000e |
| Residual | 13228.908 | 501 | 26.405 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 5 | Regression | 30246.951 | 5 | 6049.390 | 242.571 | .000f |
| Residual | 12469.344 | 500 | 24.939 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 6 | Regression | 30575.223 | 6 | 5095.870 | 209.441 | .000g |
| Residual | 12141.073 | 499 | 24.331 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 7 | Regression | 30848.060 | 7 | 4406.866 | 184.915 | .000h |
| Residual | 11868.236 | 498 | 23.832 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 8 | Regression | 31037.996 | 8 | 3879.749 | 165.113 | .000i |
| Residual | 11678.299 | 497 | 23.498 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 9 | Regression | 31132.708 | 9 | 3459.190 | 148.120 | .000j |
| Residual | 11583.588 | 496 | 23.354 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 10 | Regression | 31361.312 | 10 | 3136.131 | 136.714 | .000k |
| Residual | 11354.983 | 495 | 22.939 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| 11 | Regression | 31634.931 | 11 | 2875.903 | 128.206 | .000l |
| Residual | 11081.364 | 494 | 22.432 |  |  |
| Total | 42716.295 | 505 |  |  |  |
| a. Dependent Variable: medv | | | | | | |
| b. Predictors: (Constant), lstat | | | | | | |
| c. Predictors: (Constant), lstat, rm | | | | | | |
| d. Predictors: (Constant), lstat, rm, ptratio | | | | | | |
| e. Predictors: (Constant), lstat, rm, ptratio, dis | | | | | | |
| f. Predictors: (Constant), lstat, rm, ptratio, dis, nox | | | | | | |
| g. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas | | | | | | |
| h. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black | | | | | | |
| i. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn | | | | | | |
| j. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim | | | | | | |
| k. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad | | | | | | |
| l. Predictors: (Constant), lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad, tax  As we can see that in each step a new predictor variable has been added as we move forward. In the options, we mention the entry criteria for the variable which means that the significance value of the variable must be less than 0.05 in order to stay in the model. | | | | | | |

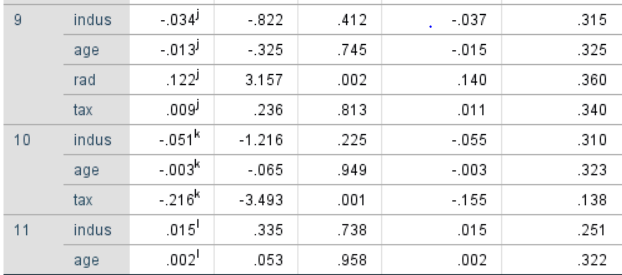
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
| B | Std. Error | Beta | Lower Bound | Upper Bound |
| 1 | (Constant) | 34.554 | .563 |  | 61.415 | .000 | 33.448 | 35.659 |
| lstat | -.950 | .039 | -.738 | -24.528 | .000 | -1.026 | -.874 |
| 2 | (Constant) | -1.358 | 3.173 |  | -.428 | .669 | -7.592 | 4.875 |
| lstat | -.642 | .044 | -.499 | -14.689 | .000 | -.728 | -.556 |
| rm | 5.095 | .444 | .389 | 11.463 | .000 | 4.222 | 5.968 |
| 3 | (Constant) | 18.567 | 3.913 |  | 4.745 | .000 | 10.879 | 26.255 |
| lstat | -.572 | .042 | -.444 | -13.540 | .000 | -.655 | -.489 |
| rm | 4.515 | .426 | .345 | 10.603 | .000 | 3.679 | 5.352 |
| ptratio | -.931 | .118 | -.219 | -7.911 | .000 | -1.162 | -.700 |
| 4 | (Constant) | 24.471 | 4.078 |  | 6.001 | .000 | 16.459 | 32.483 |
| lstat | -.665 | .047 | -.517 | -14.233 | .000 | -.757 | -.574 |
| rm | 4.224 | .424 | .323 | 9.966 | .000 | 3.391 | 5.056 |
| ptratio | -.974 | .116 | -.229 | -8.391 | .000 | -1.202 | -.746 |
| dis | -.552 | .127 | -.126 | -4.348 | .000 | -.801 | -.303 |
| 5 | (Constant) | 37.499 | 4.613 |  | 8.129 | .000 | 28.436 | 46.562 |
| lstat | -.581 | .048 | -.451 | -12.122 | .000 | -.675 | -.487 |
| rm | 4.163 | .412 | .318 | 10.104 | .000 | 3.354 | 4.973 |
| ptratio | -1.046 | .114 | -.246 | -9.212 | .000 | -1.269 | -.823 |
| dis | -1.185 | .168 | -.271 | -7.034 | .000 | -1.516 | -.854 |
| nox | -17.997 | 3.261 | -.227 | -5.519 | .000 | -24.403 | -11.590 |
| 6 | (Constant) | 36.923 | 4.559 |  | 8.099 | .000 | 27.965 | 45.880 |
| lstat | -.570 | .047 | -.442 | -12.010 | .000 | -.663 | -.477 |
| rm | 4.112 | .407 | .314 | 10.097 | .000 | 3.312 | 4.912 |
| ptratio | -1.003 | .113 | -.236 | -8.895 | .000 | -1.224 | -.781 |
| dis | -1.145 | .167 | -.262 | -6.865 | .000 | -1.472 | -.817 |
| nox | -18.740 | 3.227 | -.236 | -5.807 | .000 | -25.081 | -12.400 |
| chas | 3.244 | .883 | .090 | 3.673 | .000 | 1.509 | 4.980 |
| 7 | (Constant) | 30.412 | 4.905 |  | 6.200 | .000 | 20.774 | 40.050 |
| lstat | -.537 | .048 | -.417 | -11.204 | .000 | -.631 | -.443 |
| rm | 4.294 | .407 | .328 | 10.561 | .000 | 3.495 | 5.093 |
| ptratio | -.974 | .112 | -.229 | -8.701 | .000 | -1.194 | -.754 |
| dis | -1.123 | .165 | -.257 | -6.804 | .000 | -1.448 | -.799 |
| nox | -16.677 | 3.252 | -.210 | -5.129 | .000 | -23.066 | -10.288 |
| chas | 3.052 | .876 | .084 | 3.484 | .001 | 1.331 | 4.773 |
| black | .009 | .003 | .089 | 3.384 | .001 | .004 | .014 |
| 8 | (Constant) | 30.317 | 4.871 |  | 6.224 | .000 | 20.747 | 39.887 |
| lstat | -.543 | .048 | -.422 | -11.398 | .000 | -.637 | -.450 |
| rm | 4.116 | .409 | .314 | 10.074 | .000 | 3.313 | 4.919 |
| ptratio | -.882 | .116 | -.208 | -7.621 | .000 | -1.109 | -.654 |
| dis | -1.383 | .188 | -.317 | -7.370 | .000 | -1.751 | -1.014 |
| nox | -16.687 | 3.229 | -.210 | -5.168 | .000 | -23.031 | -10.344 |
| chas | 3.111 | .870 | .086 | 3.576 | .000 | 1.402 | 4.821 |
| black | .009 | .003 | .093 | 3.563 | .000 | .004 | .015 |
| zn | .038 | .013 | .096 | 2.843 | .005 | .012 | .064 |
| 9 | (Constant) | 29.508 | 4.873 |  | 6.056 | .000 | 19.935 | 39.081 |
| lstat | -.525 | .048 | -.408 | -10.858 | .000 | -.620 | -.430 |
| rm | 4.150 | .408 | .317 | 10.179 | .000 | 3.349 | 4.951 |
| ptratio | -.839 | .117 | -.197 | -7.147 | .000 | -1.069 | -.608 |
| dis | -1.432 | .189 | -.328 | -7.591 | .000 | -1.802 | -1.061 |
| nox | -16.089 | 3.233 | -.203 | -4.977 | .000 | -22.440 | -9.737 |
| chas | 3.030 | .868 | .084 | 3.489 | .001 | 1.324 | 4.736 |
| black | .008 | .003 | .082 | 3.084 | .002 | .003 | .014 |
| zn | .042 | .013 | .107 | 3.131 | .002 | .016 | .068 |
| crim | -.061 | .030 | -.057 | -2.014 | .045 | -.121 | -.001 |
| 10 | (Constant) | 34.712 | 5.103 |  | 6.803 | .000 | 24.687 | 44.738 |
| lstat | -.528 | .048 | -.410 | -11.019 | .000 | -.622 | -.434 |
| rm | 3.977 | .408 | .304 | 9.754 | .000 | 3.176 | 4.778 |
| ptratio | -1.015 | .129 | -.239 | -7.867 | .000 | -1.268 | -.761 |
| dis | -1.429 | .187 | -.327 | -7.647 | .000 | -1.797 | -1.062 |
| nox | -20.314 | 3.472 | -.256 | -5.850 | .000 | -27.137 | -13.492 |
| chas | 2.968 | .861 | .082 | 3.448 | .001 | 1.277 | 4.659 |
| black | .010 | .003 | .096 | 3.591 | .000 | .004 | .015 |
| zn | .037 | .013 | .093 | 2.731 | .007 | .010 | .063 |
| crim | -.105 | .033 | -.098 | -3.164 | .002 | -.170 | -.040 |
| rad | .129 | .041 | .122 | 3.157 | .002 | .049 | .209 |
| 11 | (Constant) | 36.341 | 5.067 |  | 7.171 | .000 | 26.385 | 46.298 |
| lstat | -.523 | .047 | -.406 | -11.019 | .000 | -.616 | -.429 |
| rm | 3.802 | .406 | .290 | 9.356 | .000 | 3.003 | 4.600 |
| ptratio | -.947 | .129 | -.223 | -7.334 | .000 | -1.200 | -.693 |
| dis | -1.493 | .186 | -.342 | -8.037 | .000 | -1.858 | -1.128 |
| nox | -17.376 | 3.535 | -.219 | -4.915 | .000 | -24.322 | -10.430 |
| chas | 2.719 | .854 | .075 | 3.183 | .002 | 1.040 | 4.397 |
| black | .009 | .003 | .092 | 3.475 | .001 | .004 | .015 |
| zn | .046 | .014 | .116 | 3.390 | .001 | .019 | .072 |
| crim | -.108 | .033 | -.101 | -3.307 | .001 | -.173 | -.044 |
| rad | .300 | .063 | .284 | 4.726 | .000 | .175 | .424 |
| tax | -.012 | .003 | -.216 | -3.493 | .001 | -.018 | -.005 |
| a. Dependent Variable: medv | | | | | | | | |

The above table tells us about the Beta coefficients for the predictors in each step.

The final equation after step 11 is:

Medv = 36.341 – 0.523\*lstat + 3.802\*rm – 0.947\*ptratio – 1.493\*dis – 17.376\*nox + 2.719\*chas + 0.009\*black + 0.046\*zn – 0.108\*crim + 0.3\*rad – 0.012\*tax

The table below shows the variables which were excluded in the final three steps. In the last step, we can see that indus and age variables have significance values greater than 0.05 and hence they have been not been included.



**Problem 2 (Principal Component Analysis - 20 points):** The data given in the file ‘problem3.txt’[[1]](#footnote-1) (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:

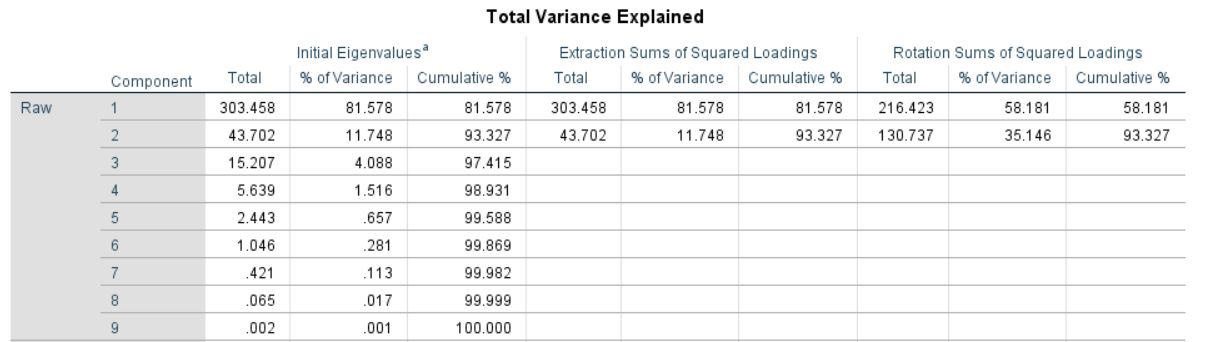
Variable Names:

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
10. TC: Percentage employed in transport and communications.

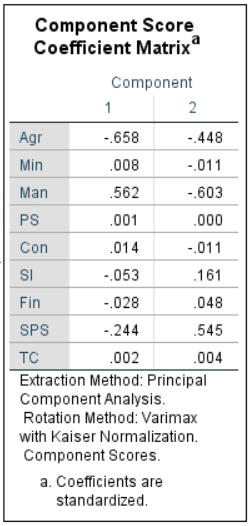
Perform a principal component analysis using the covariance matrix:

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

2 is the required number of PC since we must capture 90% of variance



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



From the above table, we can get the Equations for the two components:

**First Component:**

F1 = - 0.658\*Agr + 0.008\*Min + 0.562\*Man + 0.001\*PS + 0.014\*Con – 0.053\*SI – 0.028\*Fin – 0.244\*SPS + 0.002\*TC

**Second Component:**

F2 = - 0.448\*Agr – 0.011\*Min – 0.603\*Man + 0.000\*PS – 0.011\*Con + 0.161\*SI + 0.048\*Fin + 0.545\*SPS + 0.004\*TC

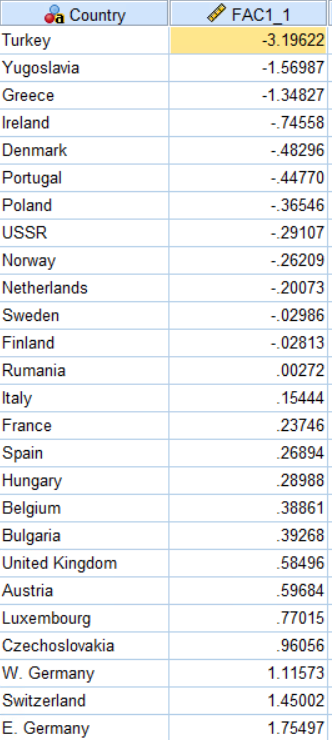
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Rotated Component Matrixa** | | | | |
|  | Raw | | Rescaled | |
| Component | | Component | |
| 1 | 2 | 1 | 2 |
| Agr | -12.693 | -8.955 | -.816 | -.576 |
| Min | .326 | -.517 | .336 | -.533 |
| Man | 6.801 | -1.417 | .970 | -.202 |
| PS | .165 | .020 | .439 | .053 |
| Con | .968 | .128 | .589 | .078 |
| SI | 1.827 | 3.198 | .399 | .699 |
| Fin | -.056 | 1.015 | -.020 | .362 |
| SPS | 2.065 | 6.061 | .302 | .887 |
| TC | .607 | .511 | .436 | .367 |
| Extraction Method: Principal Component Analysis.  Rotation Method: Varimax with Kaiser Normalization. | | | | |
| a. Rotation converged in 3 iterations. | | | | |

From the above table, we can see that Component 1 has strong correlation with Agr, Man, PS, Con and TC industries.

Component 2 has strong correlation with Min, SI, Fin and SPS industries.

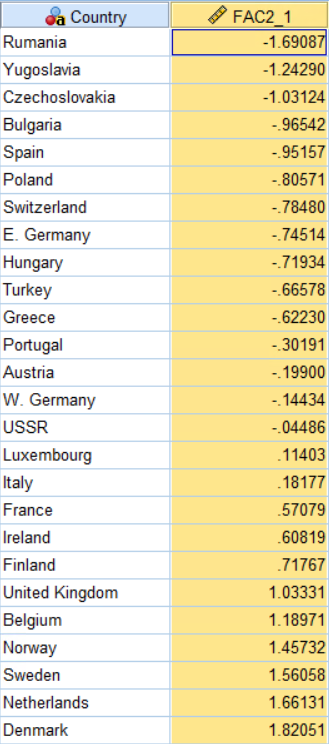
1. What countries have the highest and lowest values for each principal component (only include the number of components specified in part a). For each of those countries, give the principal component scores (again only for the number of components specified in part a).

For Component 1:

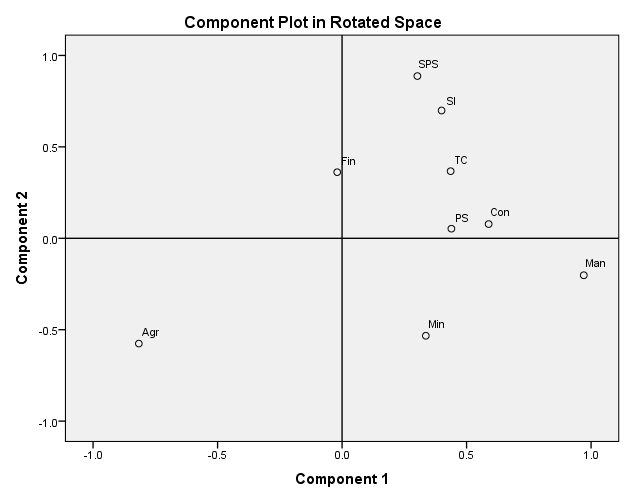


For the first component E Germany has the highest score which means that in E Germany the industries like Agr, Man, PS, Con and TC industries are the largest employers and Turkey has the least score and has less employment in these industries.

**For Component 2:**

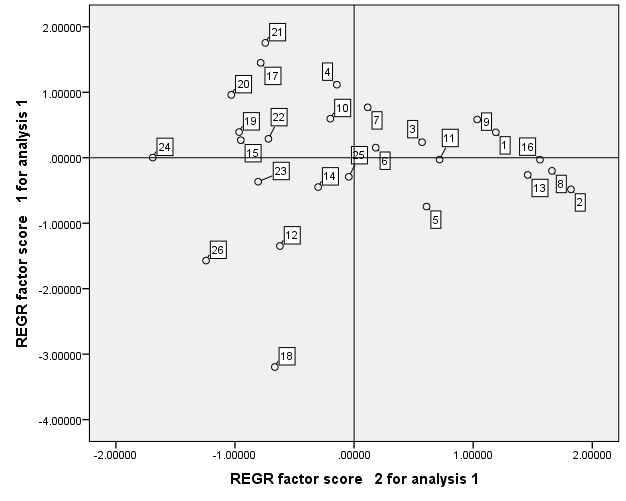


In this component Denmark has the highest score which means Min, SI, Fin and SPS industries are the largest employers and Rumania has the least score meaning these industries aren’t as large employers in this country when compared to other countries.



From the above plot, we can say that Man, Min, Con, PS, TC, SI and SPS all load positively on component 1 and Agr loads negatively on component 2 and Fin loads near to 0 in component 1 and positively on component 2.

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



As we can see that there is no relation between the two Components and this could be because the component s are orthogonal to each other.

The countries in the bottom left quadrant are under developed countries since they have all negative loadings for both the components.

The countries in the top right quadrant are more Balanced countries in which most of the industries are present and employ a good deal of people.

**Problem 3 (Extra Credit – 5 points):** Briefly describe the similarities and differences between linear regression and principal component analysis.

In Linear Regression, we have a Dependent variable and one or more independent variables. Regression equations are generated by learning through the training set and hence linear regression is a form of supervised learning. In linear regression if two or more independent variables are correlated the output is affected. In linear regression, the best fit line through the data points tries to minimize the distance between points and the line from the perspective of the axis we are regressing with respect to.

Principal Component Analysis: This is used to reduce the number of features present in a dataset. In this the closely correlated variables are merged together into a component. This merging decreases the dimensionality of the dataset and increases the explanatory power of the model. In PCA, the orthogonal distance to the model line is minimized by fitting a line through the data points in the normality plot.

**Notes on putting together the assignment:**

1. Include the SPSS output tables that support your reporting and interpretation.
2. When you write the solution for each problem, you need to break it up by the parts of the question instead of one big paragraph.

1. <http://lib.stat.cmu.edu/DASL/Datafiles/EuropeanJobs.html> [↑](#footnote-ref-1)