## Problem 1

(A)

* Goodness of fit
  + R-squared: 0.7406
  + Adj R-squared: 0.7338
* Utility of model
  + F-Statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16
* Estimated coefficients
  + (Intercept)
    - 3.646e+01
    - 5.103e+00 (std err)
    - 3.28e-12 (signif)
  + CRIM
    - -1.080e-01
    - 3.286e-02 (std err)
    - 0.001087 (signif)
  + ZN
    - 4.642e-02
    - 1.373e-02 (std err)
    - 0.000778 (signif)
  + INDUS
    - 2.056e-02
    - 6.150e-02 (std err)
    - 0.738288 (signif)
  + CHAS
    - 2.687e+00
    - 8.616e-01 (std err)
    - 0.001925 (signif)
  + NOX
    - -1.777e+01
    - 3.820e+00 (std err)
    - 4.25e-06 (signif)
  + RM
    - 3.810e+00
    - 4.179e-01 (std err)
    - < 2e-16 (signif)
  + AGE
    - 6.922e-04
    - 1.321e-02 (std err)
    - 0.958229 (signif)
  + DIS
    - -1.476e+00
    - 1.995e-01 (std err)
    - 6.01e-13 (signif)
  + RAD
    - 3.060e-01
    - 6.635e-02 (std err)
    - 5.07e-06 (signif)
  + TAX
    - -1.233e-02
    - 3.760e-03 (std err)
    - 0.001112 (signif)
  + PTRATIO
    - -9.527e-01
    - 1.308e-01 (std err)
    - 1.31e-12 (signif)
  + B
    - 9.312e-03
    - 2.686e-03 (std err)
    - 0.000573 (signif)
  + LSTAT
    - -5.248e-01
    - 5.072e-02 (std err)
    - < 2e-16 (signif)
* Evaluation
  + Good fit and model utility. Model is able to explain ~ 73% of the variation in the dependent variable.
  + Large number of highly significant coefficients.
  + Next steps would likely be:
    - remove 'AGE' (very low significance) and re-run
    - assuming 'INDUS' significance remains low following this adjustment, remove it as well, and re-run

(B)

* Order of variable addition as the result of forward selection:
  + LSTAT
  + RM
  + PTRATIO
  + DIS
  + NOX
  + CHAS
  + B
  + ZN
  + CRIM
  + RAD
  + TAX
  + (further confirmation that AGE and INDUS don’t have a significant impact on the model)

## Problem 2

* 1. Test of H0 that the canonical correlations are all equal to zero:
     + Test statistic: **.69630**
     + DF: **15.00** & **433.81**
     + P-value: **.000**
  2. Test of H0 that the second and third canonical correlations equal zero:
     + Test statistic: **.81790**
     + DF: **8.00** & **316.00**
     + P-value: **.000**
  3. Test of H0 that the third canonical correlation equals zero:
     + Test statistic: **.92841**
     + DF: **3.00** & **159.00**
     + P-value: **.008**
  4. Present the three canonical correlations, together with their standard errors.
     1. CC1: **.38558**
     2. CC2: **.34500**
     3. CC3: **.26757**
  5. What can you conclude from the above analyses?

**All are significant, given that the largest P-value is .008.**

* 1. Give the formulae for the significant canonical variates for the soil and water variables.

**F1 = .73743(THGSDFC) - .95497(TCSDFB) -.55554 (TPRSDFB)**

**F2 = -.65693(THGSDFC) - .46446(TCSDFB) +.42444(TPRSDFB)**

**F3 = -.91123(THGSDFC) + .89625(TCSDFB) -.79002(TPRSDFB)**

**G1 = .32611(MEHGSWB) + .16111(TURB) – 1.05314(DOCSWD) - .10646(SRPRSWFB) + .26750(THGFSFC)**

**G2 = -.27757(MEHGSWB) + .04268(TURB) – .39118(DOCSWD) + .51898(SRPRSWFB) - .63876(THGFSFC)**

**G3 = .20041(MEHGSWB) + .50364(TURB) – .32826(DOCSWD) - .65958(SRPRSWFB) - .61571(THGFSFC)**

* 1. 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.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **F1** | **F2** | **F3** |
| **THGSDFC** | **-.00951** | **-.88365** | **-.46806** |
| **TCSDFB** | **-.63909** | **-.76826** | **.03666** |
| **TPRSDFB** | **-.71407** | **.14767** | **-.68433** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **G1** | **G2** | **G3** |
| **MEHGSWB** | **-.21383** | **-.54424** | **.05581** |
| **TURB** | **-.12070** | **-.03436** | **.49853** |
| **DOCSWD** | **-.89202** | **-.39006** | **.02465** |
| **SRPRSWFB** | **-.17194** | **.58138** | **-.63984** |
| **THGFSFC** | **.49143** | **-.63010** | **-.52590** |

* 1. What can you conclude from the above analyses?

## Problem 3

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

**2 components required**

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

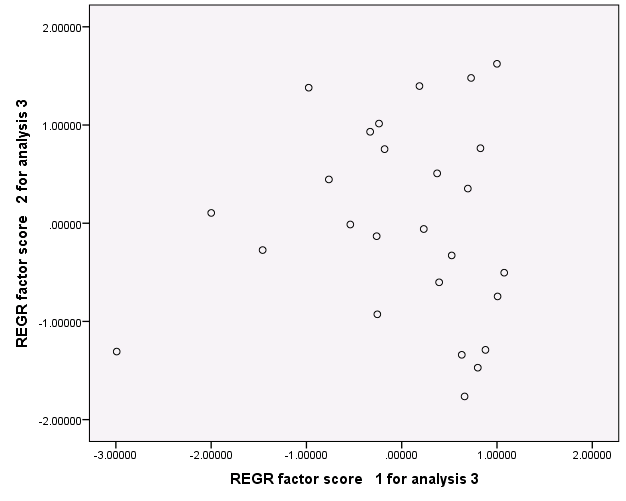
**PC1 = -.999(Agr) -.035(Min) + .674(Man) + .388(PS) + .525(Con) + .730(SI) + .193(Fin) + .760(SPS) + .568(TC)**

**{heavily service and manufacturing based / low agriculture economy}**

**PC2 = -.003(Agr) +.629(Min) + .727(Man) + .211(PS) + .277(Con) - .339(SI) - .306(Fin) - .549(SPS) - .047(TC)**

**{heavily manufacturing and industry / low service based economy}**

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).
   * PC1:
     + Hi: **UK (PC1: 1.075/PC2: -.504) and Belgium (PC1: 1.006/PC2: -.745)**
     + Low: **Turkey (PC1:-2.992/PC2:-1.307) and Yugoslavia (PC1:-2.000/PC2:.105)**
   * PC2:
     + Hi: **E. Germany (PC2:1.623/PC1:1.000) and Switzerland (PC2:1.479/PC1:.728)**
     + Low: **Denmark (PC2:-1.76/PC1:.660)and Netherlands (PC2:-1.47/PC1:.798)**
2. Include and interpret the scatter plot of the data using the first two principal components.



**The scatterplot shows no clear pattern, which is expected. This indicates that there is no correlation between the two components – part of the ‘goal’ of the PCA factor extraction.**

## Problem 4

Briefly describe the similarities and differences between:

1. Linear regression and canonical correlation
2. Canonical correlation and principal component analysis