Department of Statistics

COURSE STATS 330/762

Model answers to Assignment 2, 2011

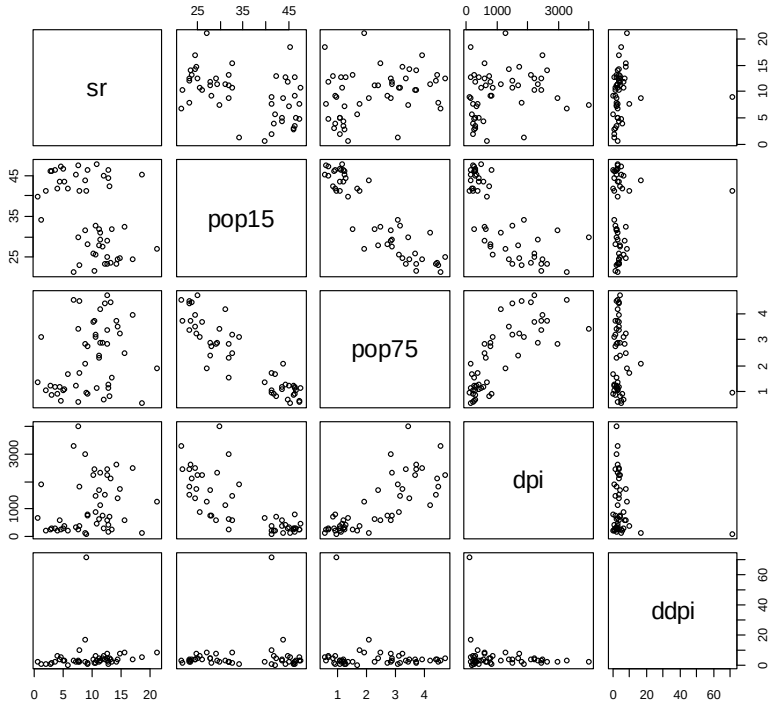
The data set for this assignment was in the file **savings.txt** which is available on the course web page.

Questions and tasks

1. *Load the data into R, and make a data frame **savings.df** to contain the data. Check for any typographical errors (the data below may be taken to be the correct data, but the data on the web may have been corrupted). Correct as necessary. Print out the last 10 lines of the data file. [5 marks]*

The following code reads in the data and draws a pairs plot:

```
infile =
"http://www.stat.auckland.ac.nz/~lee/330/datasets.dir/savings.txt"
savings.df = read.table(infile, header=TRUE)
pairs(savings.df)
```



The pairs plot indicates a point with a large value of ddpi. Examining the data reveals that this large value is the value 71.54 for India. Checking the data sheet we see this is an error, the value should be 1.54. India is the 20th row in the data:

```
> savings.df[20,]
      sr pop15 pop75   dpi  ddpi
India  9 41.31  0.96 88.94 71.54
```

We correct it as follows

```
> savings.df[20,5]=1.54
> savings.df[20,]
      sr pop15 pop75   dpi ddpi
India  9 41.31  0.96 88.94 1.54
```

There are no other errors. Finally we print out the last 10 lines:

```
> savings.df[41:50,]
                sr pop15 pop75     dpi  ddpi
Turkey        5.13 43.42  1.08  389.66  2.96
Tunisia       2.81 46.12  1.21  249.87  1.13
UnitedKingdom 7.81 23.27  4.46 1813.93  2.01
UnitedStates  7.56 29.81  3.43 4001.89  2.45
Venezuela     9.22 46.40  0.90  813.39  0.53
Zambia       18.56 45.25  0.56  138.33  5.14
Jamaica       7.72 41.12  1.73  380.47 10.23
Uruguay       9.24 28.13  2.72  766.54  1.88
Libya         8.89 43.69  2.07  123.58 16.71
Malaysia      4.71 47.20  0.66  242.69  5.08
```

[5 marks: 1 for reading in, 1 for a graph, 2 for identifying and correcting the mistake, 1 for printing out.]

2. *Fit a regression model to these data, using sr as the response. Do all variables appear to be necessary in the model? Do you think that the model using ddpi alone is an adequate submodel? Give reasons.[10 marks]*

We fit the model and produce the summary:

```
> savings.lm = lm(sr ~ pop15 + pop75 + dpi + ddpi,
data=savings.df)
> summary(savings.lm)

Call:
lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = savings.df)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75       -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi         0.4096949  0.1961971   2.088 0.042471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared: 0.3385,     Adjusted R-squared: 0.2797
F-statistic: 5.756 on 4 and 45 DF,  p-value: 0.0007904
```

Two of the variables (pop15 and ddpi) are significant at the 5% level. It seems that either pop75 or dpi could be dropped (but we can't tell from this output if both could be dropped.
[3 marks for identifying the two variables that seem insignificant, 2 marks for saying they can't both be dropped on this evidence]

To see if the model with ddpi alone is adequate, we use the code

```
> ddpi.lm = lm(sr ~ ddpi, data=savings.df)
> anova(ddpi.lm, savings.lm)
Analysis of Variance Table

Model 1: sr ~ ddpi
Model 2: sr ~ pop15 + pop75 + dpi + ddpi
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     48 892.25
2     45 650.71  3    241.54 5.5679 0.002456 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since the p-value is very small, we conclude that the model with ddpi alone is not adequate.
[2 marks for the output, 3 marks for the correct conclusion. A total of 10 marks for the question]

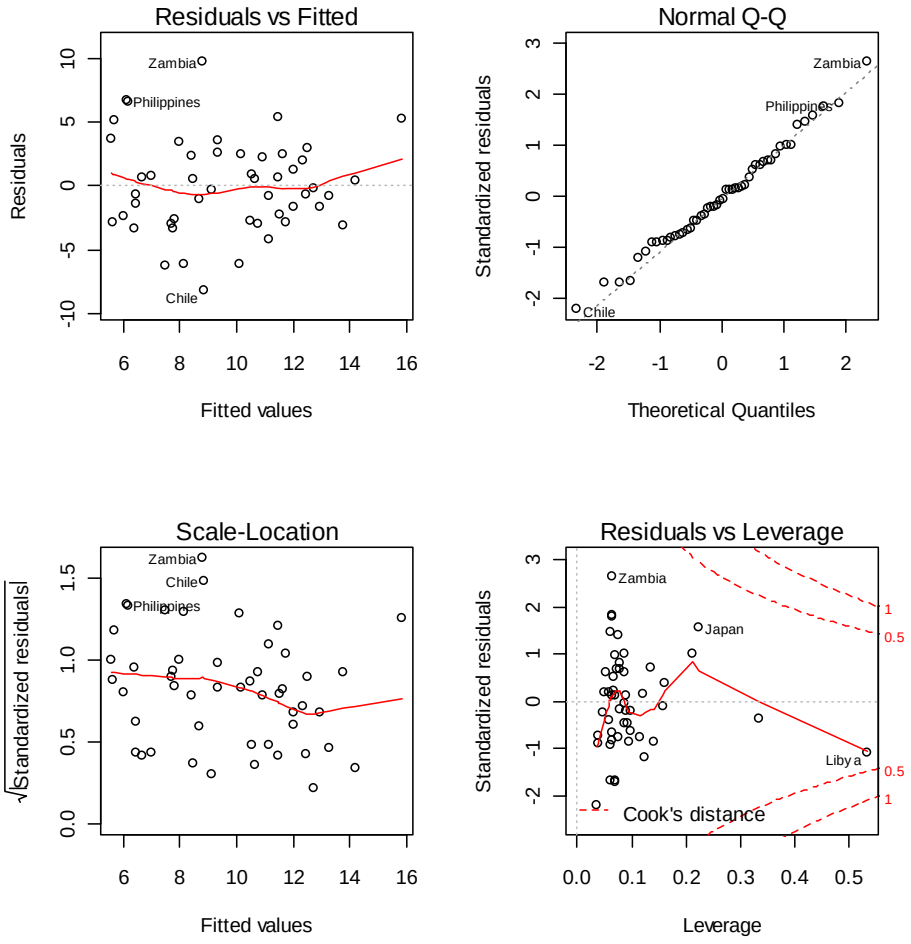
3. Do you think that the data support the life-cycle savings hypothesis? Give reasons. [5 marks]

On the face of it, the support for the theory is rather weak. Only two of the variables are significant, and the R^2 is not very convincing. Still there is some relationship between the explanatory variables and the savings rate, as the hypothesis of no relationship is soundly rejected (p= 0.0007904)
[2 marks for commenting on the p-values in the regression summary, 1 for the R^2, and 2 for commenting on the overall significance of the regression, 5 in total for Q3]

4. Are there any data points that unduly influence the results? What aspect of the results are influenced by what data points? Does this make you change your mind about the answer to Q3? Give a full discussion with reasons. [15 marks]

We first examine diagnostic plots:

```
> par(mfrow=c(2,2))
> plot(savings.lm)
```



[4 marks for basic residual plots]

Zambia and Chile have a rather large residuals, and Libya sees quite influential. We can compute "leave-one-out" diagnostics to get a better idea of how these points are affecting the fit:

```
>influence.measures(savings.df)
```

Only the 3 countries (Chile, Zambia and Libya) mentioned above plus the USA have diagnostics above the threshhold. They are shown below:

```
               dfb.1_  dfb.pp15 dfb.pp75  dfb.dpi  dfb.ddpi   dffit cov.r   cook.d    hat inf
Chile         -0.19941  0.132652  0.21979 -0.01998  0.120007 -0.4554 0.655 3.78e-02 0.0373   *
UnitedStates   0.06910 -0.072886  0.03745 -0.23312 -0.032729 -0.2510 1.655 1.28e-02 0.3337   *
Zambia         0.16361 -0.079172 -0.33899  0.09406  0.228232  0.7482 0.512 9.66e-02 0.0643   *
Libya          0.55074 -0.483244 -0.37974 -0.01937 -1.024477 -1.1601 2.091 2.68e-01 0.5315   *
```

All these are having an effect on the standard errors (cov.r more than 0.3 = 3p/n away from 1), USA and Libya have high leverage, Libya has an effect on the coefficient of ddpi, and its own fitted value.

[2 marks for computing influence measures and/or influence plots, 4 marks for sensible comments on them]

We can refit leaving out Libya:

```
>savings.49.lm = lm(sr ~ pop15 + pop75 + dpi + ddpi, data=savings.df,
subset=-49)
> summary(savings.49.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.5240460  8.2240263   2.982  0.00465 **
pop15       -0.3914401  0.1579095  -2.479  0.01708 *
pop75       -1.2808669  1.1451821  -1.118  0.26943
dpi         -0.0003189  0.0009293  -0.343  0.73312
ddpi         0.6102790  0.2687784   2.271  0.02812 *

Residual standard error: 3.795 on 44 degrees of freedom
Multiple R-squared: 0.3554,     Adjusted R-squared: 0.2968
F-statistic: 6.065 on 4 and 44 DF,  p-value: 0.0005617
```

The R^2 has improved a bit but the variable pop15 is not as significant as before. The conclusions have not really changed.

[2 marks for refitting, 3 marks for commenting that no real change in the conclusions is called for. Total marks for Q4 is 15]

5. *Do you think these data have collinearity problems? Give reasons. What is the consequence if they do? [5 marks]*

We calculate the VIF's, using the data with Libya excluded.

```
> VIFs = diag(solve(cor(savings.df[-49, -1])))
> VIFs
   pop15    pop75      dpi     ddpi
6.975529 7.430085 2.826126 1.165772
```

The VIF's for pop15 and pop75 are quite high, indicating that these variables are not being estimated very well in the full model, even when Libya is left out.

[2 marks for computing VIF's (either with or without Libya), 3 marks for comments. 5 marks total fro Q5]

Total marks are 40.