ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Applied Biostatistics - MATH-493

Assignment 1: Multiple Regression in R, Life-Cycle Savings

Authors:

Nils Kalbfuss, Jules Mertens,
Yan Wong Wen,

Professor:

Darlene Goldstein

April 11, 2018

Contents

1 Introduction

In order to predict the savings ratio (aggregate personal savings divided by disposable income) for any country at any time, 5 variables were measured in 50 different countries: the numeric aggregate personal savings (sr), the per-capita disposable income (dpi), the percentage rate of change in per-capita disposable income (ddpi), and two demographic variables: the percentage of population less than 15 years old (pop15) and the percentage of the population over 75 years old (pop75). Since the business cycle might have an influence on savings, data were averaged over a whole decade (1960-1970). Here, we aim at establishing a model that is able to predict the personal savings ratio based on the other four recorded observations, by means of multiple regression.

2 Data exploration

A first step in the statistical analysis is an exploration of the dataset, to get a first idea of important relations between the variables which will be useful in the modelling phase. The bivariate relations between each pair of variables are explored by means of a scatter-plot matrix (Fig.1). The two demographic variables pop15 and pop75 show a strong negative trend, which looks quite linear. The pairs of variables (pop15, dpi) and (pop75, dpi) show a less strong negative, respectively positive trend.

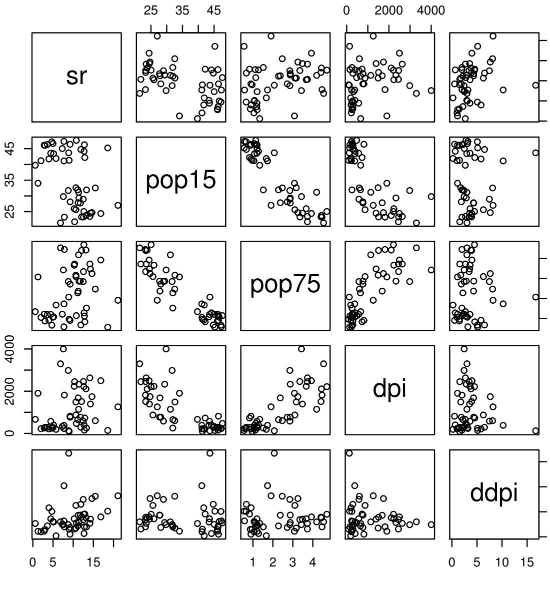

Figure 1: Scatter-plot matrix of bivariate relations between the variables

These three trends are also confirmed by the correlation matrix (Table 1). Most striking correlations are a negative correlation between pop15 and pop75 of -0.908, a positive correlation between pop75 and dpi of 0.787 and a negative correlation between pop15 and dpi of -0.756. The analysis of partial correlations (Table 2) revealed that pop15 and pop75 are indeed strongly correlated (ρ = -0.771). However, the relationship between dpi and the demographic variables is mostly due to underlying relations: partial correlation between pop15 and dpi dropped to ρ = -0.185 and partial correlation between pop75 and dpi dropped to ρ = 0.345.

	sr	pop15	pop75	dpi	ddpi
sr	1.00	-0.46	0.32	0.22	0.30
pop15	-0.46	1.00	-0.91	-0.76	-0.05
pop75	0.32	-0.91	1.00	0.79	0.03
dpi	0.22	-0.76	0.79	1.00	-0.13
ddpi	0.30	-0.05	0.03	-0.13	1.00

Table 1: Correlation Matrix

	sr	pop15	pop75	dpi	ddpi
sr	1.00	-0.43	-0.23	-0.05	0.30
pop15	-0.43	1.00	-0.77	-0.19	0.04
pop75	-0.23	-0.77	1.00	0.34	0.12
dpi	-0.05	-0.19	0.34	1.00	-0.23
ddpi	0.30	0.04	0.12	-0.23	1.00

Table 2: Partial Correlation Matrix

An important remark here is that the correlation matrices only consider linear association between the variables. The data exploration showed that a strong linear association exists only between the demographic variables. Other kinds of non-linear associations can possibly exist between the other variables but are not obvious at first sight.

The variable of interest, the savings ratio sr, shows no clear trend with any of the other explanatory variables. The first column of the partial correlation matrix (Table 2) shows that variables pop15 and ddpi have the strongest linear association with sr. Therefore they are probably the most important explanatory variables in a linear model to predict sr.

3 Establishing an additive linear model

According to the life-cycle savings hypothesis, developed by Franco Modigliani, the savings ratio is explained by all 4 variables. Therefore, we started modelling using an additive linear model that includes all variables. In this model only pop15 (p=0.0026) and ddpi were significant (p=0.0425) under a significance level of $\alpha = 0.05$ (Table 3). The adjusted R^2 for this model is 0.2797.

Model 1: sr $\sim$ pop15 + pop75 + dpi + ddpi

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.5661	7.3545	3.88	0.0003
pop15	-0.4612	0.1446	-3.19	0.0026
pop75	-1.6915	1.0836	-1.56	0.1255
dpi	-0.0003	0.0009	-0.36	0.7192
ddpi	0.4097	0.1962	2.09	0.0425

Table 3: Coefficient estimates of the full additive model

4 Improvements of the model

In order to improve the model by exploring nested models, forward selection and backward elimination are performed. These methods use the Akaike Information criterion (or AIC-value) to compare the relative quality of different nested models. Both methods excluded dpi from the full additive model, which leads to an increase of the adjusted R^2 of 4.86% to 0.2933. Since the savings ratio (sr) is already normalized by the disposable income (dpi), it is indeed possible that this variable doesn't play an important role in predicting the savings ratio itself.

Model 2: $\quad$ sr ∼ pop15 + ddpi + pop75

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.1247	7.1838	3.92	0.0003
pop15	-0.4518	0.1409	-3.21	0.0025
ddpi	0.4278	0.1879	2.28	0.0275
pop75	-1.8354	0.9984	-1.84	0.0725

Table 4: Coefficients of the simplified model obtained by forward selection and backwards elimination

5 Multicollinearity

If two explanatory variables are highly correlated (referred to as multicollinearity) the addition of the second variable will not add a lot of valuable information to the model and should be avoided since it will result in a less robust and less stable solution. Pop15 and pop75, as already observed in the scatter plot, are very likely candidates because of their rather high linear correlation. Indeed, a VIF bigger than five (Table 5) is observed for both pop15 and pop75 which is a strong indicator for multicollinearity between those two variables.

Based on this observation, we performed backward elimination on model 1, once without pop15 and once without pop75. Both times the variable dpi was removed from the model. Since the resulting model sr ∼ pop15 + ddpi has a higher adjusted R^2 than sr ∼ pop75 + ddpi, it is preferable to eliminate pop75 and not pop15. So the resulting model is:

model 3: $\quad$ sr ∼ pop15 + ddpi

	pop15	pop75	ddpi
VIF:	5.745478	5.736014	1.004186

Table 5: Detection of multicollinearity with VIF on model 2

6 Regression diagnostics

Furthermore, we validated the assumptions about the dataset and analysed model 3 regarding outliers and influential points. In the residuals vs fitted plot (Fig. 2), all data points are randomly distributed around 0. This shows that the assumption of homoscedasticity is fulfilled. Furthermore Zambia, Chile and Japan have very high (absolute) residuals. These are therefore possible outliers, an extra control on the validity of these data points is recommended. In addition, The QQ plot shows that there is an underprediction at both tails. Again, Chile, Japan and Zambia appear here.

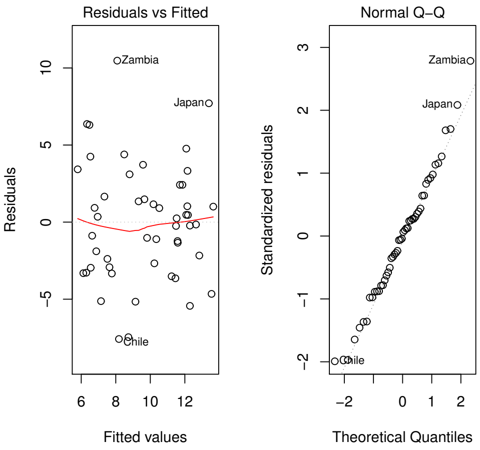

Figure 2: regression diagnostics on model 3

Finally, the Cook's distance plot and the residuals vs leverage plot (Fig. 3) show that although Zambia, Japan and Chile have the highest residuals, it is Libya that has the most influence on the model (Cook's distance of almost 0.8) by a combination of a quite high residual with a very high leverage. It is therefore very important to check the validity of this data point.

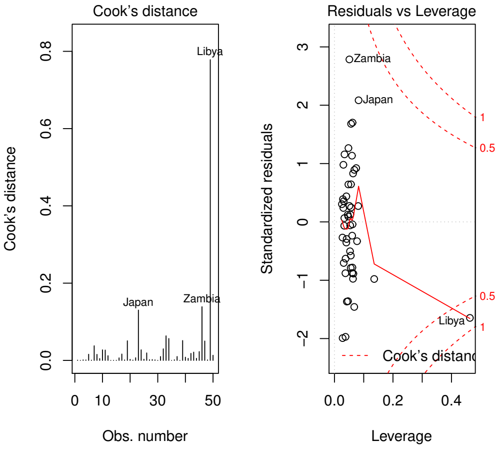

Figure 3: regression diagnostics on model 3

When taking a closer look at the data of Libya, it has a much higher ddpi (of almost 17%) than all other countries (Fig. 4). This could be a reason for non-validity of our model for this country. When excluding this data-point, the adjusted R^2 increases with 16.2% from 0.2575 to 0.2992 which illustrates its big influence on the model.

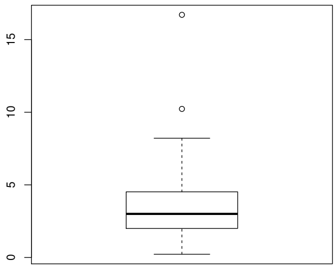

Figure 4: box plot of the ddpi (rate of change of disposable income per capita)

7 Including interactions in modelling approach

In a parallel attempt we also included interactions into our modelling approach. Inclusion of all variables and interactions lead to a very poor result with no significant predictor. We used backward elimination (comparison of nested models by means of AIC-comparison) to get rid of predictors and interactions that are less important. This finally lead to the model:

$$\text{sr} \sim \text{pop15} + \text{dpi} + \text{dpi:ddpi}$$

Although this model reached an adjusted R^2 of 0.3332, higher than the other explored models, we declined it because of its unclear economic interpretation.

8 Discussion

The final model we would propose would therefore be the following:

$$\text{Savings ratio y} = 15.15557 \text{ - } 0.20835 \text{ pop15} + 0.45342 \text{ ddpi} + \text{e}$$

In this model, pop15 appeared to be the strongest predictor (p= 0.000309). This shows that the savings ratio is negatively associated with the population less than 15 years old (pop15) and positively associated with percentage rate of change in per-capita disposable income (ddpi). The explained variance of the dataset, measured by the adjusted R^2, remains quite low which is partly due to the simplicity of our linear model.

Model	Adjusted R^2
sr ~ pop15 + pop75 + dpi + ddpi	0.2797
sr ~ pop15 + pop75 + ddpi	0.2933
sr ~ pop75 + ddpi	0.1538
sr ~ pop15 + ddpi	0.2575
sr ~ pop15 + ddpi (dropping Libya)	0.2992
sr ~ pop15 + dpi + dpi:ddpi	0.3332

Table 6: Overview: explored models and their adjusted R^2 values