Israel Diego

3/10/16

STATS 500 Homework 4

1. Compute and comment on the condition numbers.

To compute the condition numbers we first calculate the eigenvalues of the (X’X) matrix. The condition numbers correspond to the square root of the ratio of the maximum eigenvalue and the rest of the other eigenvalues.

Running the following command: > round(sqrt(max(e$val)/e$val), 3)

[1] 1.000 17.855 25.153 60.785 1647.478 5751.216

The last three condition numbers are above 30, which is considered large. We notice a wide range of condition numbers, indicating the presence of multicollinearity in more than one variable.

2. Compute and comment on the correlations between the predictors.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Longley Data Correlation Matrix | | | | | | | |
|  | GNP.deflator | GNP | Unemployed | Armed.Forces | Population | Year | Employed |
| GNP.deflator | 1 | 0.99 | 0.62 | 0.46 | 0.98 | 0.99 | 0.97 |
| GNP | 0.99 | 1 | 0.6 | 0.45 | 0.99 | 1 | 0.98 |
| Unemployed | 0.62 | 0.6 | 1 | -0.18 | 0.69 | 0.67 | 0.5 |
| Armed.Forces | 0.46 | 0.45 | -0.18 | 1 | 0.36 | 0.42 | 0.46 |
| Population | 0.98 | 0.99 | 0.69 | 0.36 | 1 | 0.99 | 0.96 |
| Year | 0.99 | 1 | 0.67 | 0.42 | 0.99 | 1 | 0.97 |
| Employed | 0.97 | 0.98 | 0.5 | 0.46 | 0.96 | 0.97 | 1 |

The above table demonstrates the correlations between predictors and the response. We see that GNP.deflator, GNP, Population, and Year are all heavily correlated with each other, which have correlation coefficients close to 1. However they are not highly correlated with the variables Armed.Forces and Unemployed, and the correlation between Armed Forces and Unemployed is -0.18, which means their correlation is relatively small and the two variables are actually anti-correlated. Observing these correlations tells us that Armed.Forces and Unemployed will be the most important variables in our regression, and we should consider keeping one or two of the other four predictors, but certainly not all of them.

3. Compute and comment on the variance inflation factors.

Using the X matrix from problem 1), we compute the VIFs.

Running the following command: > round(vif(X), 3)

GNP.deflator GNP Unemployed Armed.Forces Population Year

135.532 1788.513 33.619 3.589 399.151 758.981

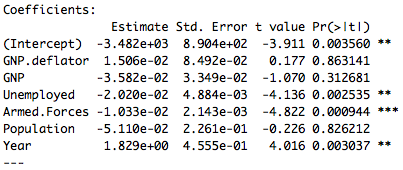
The variables with the largest variance inflations are GNP, Population, and Year. As explained in the Faraway textbook, taking the square root of the variance inflation of GNP for example, , means that the standard error of GNP is 42.2908 times higher, compared to a model without collinearity. In this case out of all the variables excluding Unemployed and Armed.Forces, GNP.deflator has the smallest VIF.

4. Choose a reduced set of predictors that does not exhibit as much collinearity as the full set, fit a new linear model with this reduced set, and comment on the differences between the reduced model and the full model.

From an economic standpoint I am inclined to use GNP exclusively as the extra variable. My intuition is that economic performance (GNP) would have some sort of relationship to number of people employed. In good times more people are employed, and in not so good times, less people will be employed. However there are two kinds of GNP variables. I will use the GNP.deflator variable because it has a significantly lower VIF than its counterpart, as observed in problem 3.

Below I show the summary statistics for the old regression model and our new regression model.

**Old Model**

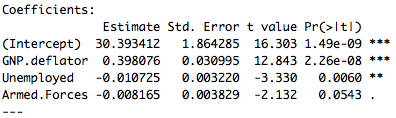


Residual standard error: 0.3049 on 9 degrees of freedom

Multiple R-squared: 0.9955, Adjusted R-squared: 0.9925

F-statistic: 330.3 on 6 and 9 DF, p-value: 4.984e-10

**New Model**



Residual standard error: 0.6776 on 12 degrees of freedom

Multiple R-squared: 0.9702, Adjusted R-squared: 0.9628

F-statistic: 130.3 on 3 and 12 DF, p-value: 2.02e-09

Comparing the two regression models, we notice a big change in the magnitude of the coefficients, but the signs of GNP.deflator, Unemployed, and Armed.Forces are still the same from the old model to the new model. The R-squared statistic has dropped a bit from the old model to the new model, but it is still high in the new model. We have gained a lot of statistical significance in our regression coefficients, with GNP.deflator and Unemployed significant at least at the 1% level, while Armed.Forces is almost significant at the 5% level. The F-statistic in the new model implies that we reject the null hypothesis that all coefficients are equal to zero. Overall the new (reduced) model is simpler and the variables now have more explanatory power in explaining the Employed variable.