|  |  |
| --- | --- |
| STA 5207 Assignment 5  Due Friday October 15 |  |

The data set Longley in the faraway package is a macroeconomic data set with 7 economical variables, observed yearly from 1947 to 1962 (*n=16*). The response variable is “Employed.”

GNP.deflator: GNP implicit price deflator (*1954=100*)

GNP: Gross National Product.

Unemployed: number of unemployed.

Armed.Forces: number of people in the armed forces.

Population: ‘noninstitutionalized’ population *≥* 14 years of age.

Year: the year (time).

Employed: number of people employed.

The data set is saved in SAS Studio as Longley.txt.

1. (10 points) How well does this model fit the data?
   1. Based on the high adj. R2 = 0.9925 and F-statistic=330.3 with p-value < 0.001, it appears that the model fits the data well.
2. (10 points) Based on the standard plots, do the constant variance and normality assumptions reasonable?
   1. These assumptions are reasonable as the histogram and boxplots are roughly bell-shaped. Also, the p-values for the normality tests are high, so fail to reject H0.
3. (10 points) We did not discuss this much, but why would it be reasonable to suspect that the errors might not be independent?
   1. The errors associated with GNP would more than likely be related to Unemployed, GNP deflator, and possibly Armed Forces.

(ACTUAL ANSWER: The data was collected each year)

1. (16 points) How many pairs of predictors are highly correlated? Consider “highly” to be above .6. Based on this, do you believe that this data set has a serious problem with multicollinearity?

GNP.deflator GNP Unemployed Armed Forces Population Year Employed

GNP.deflator 1.0000000 0.9915892 0.6206334 0.4647442 0.9791634 0.9911492 ~~0.9708985~~

GNP 0.9915892 1.0000000 0.6042609 0.4464368 0.9910901 0.9952735 ~~0.9835516~~

Unemployed 0.6206334 0.6042609 1.0000000 -0.1774206 0.6865515 0.6682566 ~~0.5024981~~

Armed.Forces 0.4647442 0.4464368 -0.1774206 1.0000000 0.3644163 0.4172451 ~~0.4573074~~

Population 0.9791634 0.9910901 0.6865515 0.3644163 1.0000000 0.9939528 ~~0.9603906~~

Year 0.9911492 0.9952735 0.6682566 0.4172451 0.9939528 1.0000000 ~~0.9713295~~

Employed ~~0.9708985 0.9835516 0.5024981 0.4573074 0.9603906 0.9713295 1.0000000~~

* 1. This data set has a massive problem with multicollinearity, where only Armed Forces appears not to be highly correlated with any other predictors.

1. (14 points) Give the variance inflation factors. Which coefficient estimates are highly inflated by the multicollinearity?

Tolerance and Variance Inflation Factor

|  |  |  |  |
| --- | --- | --- | --- |
|  | Variables | Tolerance | VIF |
| 1 | GNP.deflator | 0.0073783075 | 135.53244 |
| 2 | GNP | 0.0005591235 | 1788.51348 |
| 3 | Unemployed | 0.0297451814 | 33.61889 |
| 4 | Armed.Forces | 0.2786345641 | 3.58893 |
| 5 | Population | 0.0025053174 | 399.15102 |
| 6 | Year | 0.0013175567 | 758.98060 |

* 1. GNP, Year, and Population are highly inflated by the multicollinearity.

1. (16 points) Are there any variables that are nearly orthogonal to the others? Consider a low R2 to be less than .3.
   1. There are no variables that satisfy the condition R2 < 0.3.
2. (14 points) Give at the condition indices. How serious does the multicollinearity appear to be?

Eigenvalue and Condition Index

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Eigenvalue | Condition Index | intercept | GNP.deflator | GNP | Unemployed | Armed.Forces | Population | Year |
| 1 | 6.8614 | 1.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0004 | 0.0000 | 0.0000 |
| 2 | 0.0821 | 9.1417 | 0.0000 | 0.0000 | 0.0000 | 0.0143 | 0.0919 | 0.0000 | 0.0000 |
| 3 | 0.0457 | 12.2557 | 0.0000 | 0.0000 | 0.0003 | 0.0008 | 0.0636 | 0.0000 | 0.0000 |
| 4 | 0.0107 | 25.3366 | 0.0000 | 0.0003 | 0.0011 | 0.0646 | 0.4267 | 0.0000 | 0.0000 |
| 5 | 0.0001 | 230.4239 | 0.0000 | 0.4568 | 0.0157 | 0.0056 | 0.1154 | 0.0097 | 0.0000 |
| 6 | 0.0000 | 1048.0803 | 0.0001 | 0.5046 | 0.3284 | 0.2253 | 0.0000 | 0.8306 | 0.0002 |
| 7 | 0.0000 | 43275.0435 | 0.9999 | 0.0383 | 0.6546 | 0.6893 | 0.3020 | 0.1597 | 0.9998 |

* 1. The condition number is ~43275, so the multicollinearity is extremely serious.

1. (10 points) How many linear near-dependencies are likely causing most of the problem?
   1. 2 linear near-dependencies in condition index 6 and 7 are likely causing this problem.

Submit the following output: Correlation table, VIFs and condition numbers.