

STA4203

Assignment 6

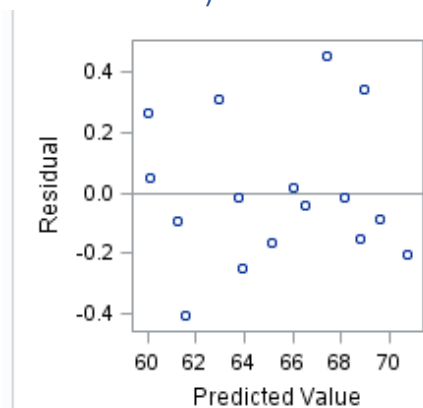
PROBLEM 1)

| Analysis of Variance | | | | | |
|----------------------|----|----------------|-------------|---------|--------|
| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
| Model | 6 | 184.17240 | 30.69540 | 330.29 | <.0001 |
| Error | 9 | 0.83642 | 0.09294 | | |
| Corrected Total | 15 | 185.00883 | | | |

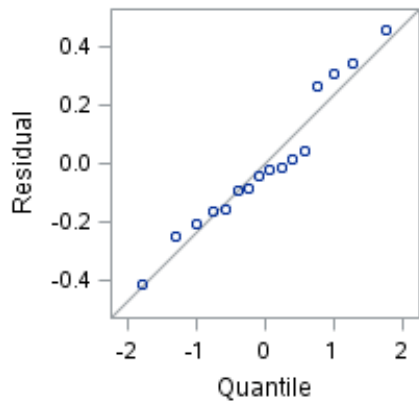
| | | | |
|----------------|----------|----------|--------|
| Root MSE | 0.30485 | R-Square | 0.9955 |
| Dependent Mean | 65.31700 | Adj R-Sq | 0.9925 |
| Coeff Var | 0.46673 | | |

Using the model with all 6 predictors, we observe that **99.55 percent** of the variance in the number of people employed can be explained by the GNP deflator, GNP, number of unemployed, number of people in the armed forces, the 'noninstitutionalized' population with more than 14 years of age and the year. At a 5% significance level, we conclude that there is a significant relationship between the predictors and the number of people employed.

PROBLEM 2)

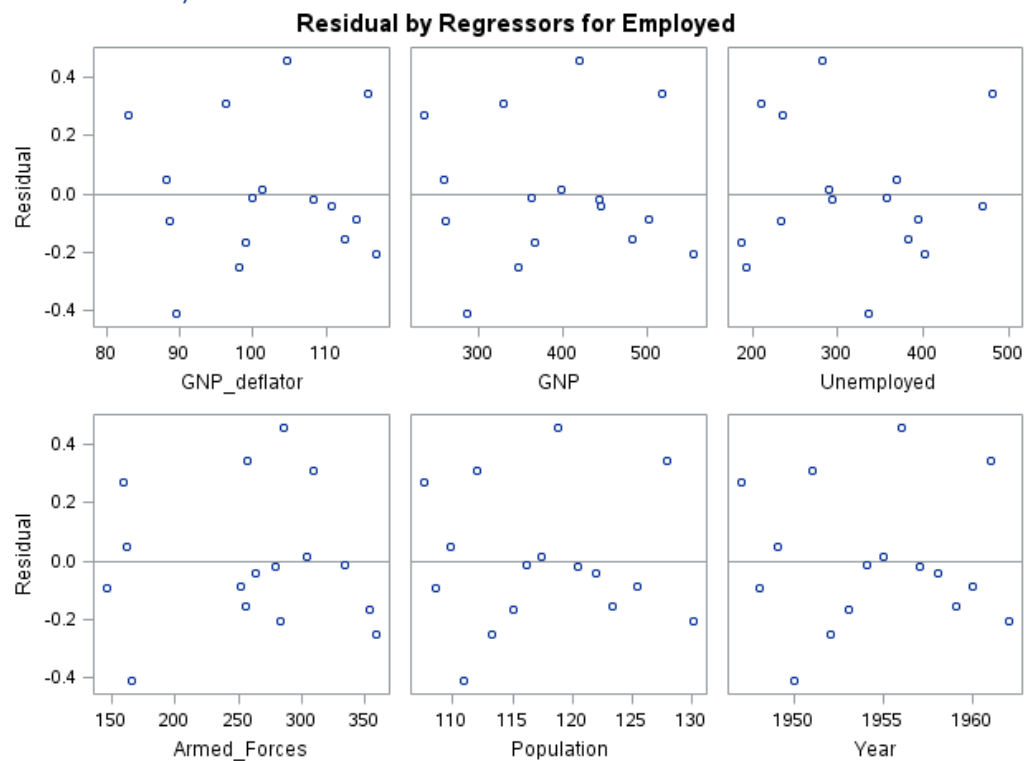


The residuals vs predicted values plot indicates that the assumption of constant variance is met since the points are scattered randomly around 0.



Analyzing the residual vs quantile plot, we can observe that the assumption of normality.

PROBLEM 3)



Since the residuals vs regressors plots don't seem to follow a pattern we can assume homoscedasticity. However, the standard error values varies across independent variables meaning that the error terms could be dependent.

PROBLEM 4)

| Correlation | | | | | | | |
|--------------|--------------|--------|------------|--------------|------------|--------|----------|
| Variable | GNP_deflator | GNP | Unemployed | Armed_Forces | Population | Year | Employed |
| GNP_deflator | 1.0000 | 0.9916 | 0.6206 | 0.4647 | 0.9792 | 0.9911 | 0.9709 |
| GNP | 0.9916 | 1.0000 | 0.6043 | 0.4464 | 0.9911 | 0.9953 | 0.9836 |
| Unemployed | 0.6206 | 0.6043 | 1.0000 | -0.1774 | 0.6866 | 0.6683 | 0.5025 |
| Armed_Forces | 0.4647 | 0.4464 | -0.1774 | 1.0000 | 0.3644 | 0.4172 | 0.4573 |
| Population | 0.9792 | 0.9911 | 0.6866 | 0.3644 | 1.0000 | 0.9940 | 0.9604 |
| Year | 0.9911 | 0.9953 | 0.6683 | 0.4172 | 0.9940 | 1.0000 | 0.9713 |
| Employed | 0.9709 | 0.9836 | 0.5025 | 0.4573 | 0.9604 | 0.9713 | 1.0000 |

From the correlation table, we can observe that the predictors **GNP_deflator**-**GNP**, **GNP_deflator**-**Unemployed**, **GNP_deflator**-**Population**, **GNP_deflator**-**Year**, **GNP**-**Unemployed**, **GNP**-**Population**, **GNP**-**Year**, **Unemployed**-**Population**, **Unemployed**-**Year** and **Population**-**Year** are highly correlated pairs of predictors because their values are all greater than 0.6. Since most of the variables are correlated, I would say this dataset has a serious problem with multicollinearity.

PROBLEM 5)

| Parameter Estimates | | | | | | |
|---------------------|----|--------------------|----------------|---------|---------|--------------------|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > t | Variance Inflation |
| Intercept | 1 | -3482.25864 | 890.42038 | -3.91 | 0.0036 | 0 |
| GNP_deflator | 1 | 0.01506 | 0.08491 | 0.18 | 0.8631 | 135.53244 |
| GNP | 1 | -0.03582 | 0.03349 | -1.07 | 0.3127 | 1788.51349 |
| Unemployed | 1 | -0.02020 | 0.00488 | -4.14 | 0.0025 | 33.61889 |
| Armed_Forces | 1 | -0.01033 | 0.00214 | -4.82 | 0.0009 | 3.58893 |
| Population | 1 | -0.05110 | 0.22607 | -0.23 | 0.8262 | 399.15102 |
| Year | 1 | 1.82915 | 0.45548 | 4.02 | 0.0030 | 758.98060 |

The VIF for **GNP_deflator**, **GNP**, **Unemployed**, **Population** and **Year** are all greater than 10. This indicates that the estimates for these predictors are highly inflated by multicollinearity.

PROBLEM 6)

The only predictor not dependent on others is **Armed_Forces**. This means it is the only predictor that is orthogonal to the others. The R² values for the predictors are all above the 0.3 threshold, **Armed_Forces** being the lowest at 0.67.

PROBLEM 7)

| Collinearity Diagnostics | | | | | | | | | |
|--------------------------|-------------|-----------------|-------------------------|--------------|-------------|------------|--------------|-------------|-------------|
| Number | Eigenvalue | Condition Index | Proportion of Variation | | | | | | |
| | | | Intercept | GNP_deflator | GNP | Unemployed | Armed_Forces | Population | Year |
| 1 | 6.86139 | 1.00000 | 1.54013E-10 | 0.00000164 | 6.742617E-7 | 0.00004472 | 0.00035369 | 1.740763E-7 | 1.54148E-10 |
| 2 | 0.08210 | 9.14172 | 8.16629E-10 | 7.095535E-9 | 0.00000753 | 0.01428 | 0.09191 | 4.021693E-8 | 7.70535E-10 |
| 3 | 0.04568 | 12.25574 | 3.342247E-8 | 1.012272E-7 | 0.00025717 | 0.00083626 | 0.06357 | 0.00000839 | 3.19652E-8 |
| 4 | 0.01069 | 25.33661 | 1.19104E-9 | 0.00034484 | 0.00107 | 0.06464 | 0.42672 | 0.00001821 | 1.426706E-9 |
| 5 | 0.00012923 | 230.42395 | 5.260203E-7 | 0.45677 | 0.01566 | 0.00559 | 0.11540 | 0.00968 | 5.273968E-7 |
| 6 | 0.00000625 | 1048.08030 | 0.00014914 | 0.50456 | 0.32839 | 0.22534 | 6.865016E-7 | 0.83056 | 0.00016031 |
| 7 | 3.663846E-9 | 43275 | 0.99985 | 0.03833 | 0.65463 | 0.68926 | 0.30205 | 0.15973 | 0.99984 |

There are three condition indices which are greater than 100. This indicates that there is **strong** collinearity.

PROBLEM 8)

The collinearity diagnostic above shows three condition indices greater than 100. This indicates that there might be 3 strong sources causing multicollinearity.

SAS CODE

```
/* Read file spider.txt and store it in dataset spiders */
FILENAME longley '/home/eff100/my_courses/jhshows0/Data Sets/longley.txt';
Data macroecon;
INFILE longley;
INPUT GNP_deflator GNP Unemployed Armed_Forces Population Year Employed;
run;

/* Problem 1-5*/
/*SSR for full model*/
PROC REG data=macroecon;
MODEL Employed= GNP_deflator GNP Unemployed Armed_Forces Population Year/vif;
plot r.*p.;
OUTPUT out=resids1 r=resid p=pred;
run;

/* Check for normality */
PROC UNIVARIATE data=resids1 normal plots;
var resid;
run;

/* Look for correlation between variables by using coefficients */
proc corr data=macroecon;
var GNP_deflator GNP Unemployed Armed_Forces Population Year;
run;
```