Title: Employment from 1947 to 1962

Data Summary

The dataset, *longley*, contains 16 observations and 7 variables. The data set contains one dependent variable, TOTEMPL (total employment) and six independent variables: INFLAT (inflation), GNP (gross national product), UNEMPL (unemployment), ARMYEMPL (size of armed forces), POPGT14 (population aged 14 and over), and YEAR (year).

```
> summary(longley)
    TOTEMPL          INFLAT            GNP             UNEMPL        ARMYEMPL   
 Min.   :60171   Min.   : 83.00   Min.   :234289   Min.   :1870   Min.   :1456  
 1st Qu.:62712   1st Qu.: 94.53   1st Qu.:317881   1st Qu.:2348   1st Qu.:2298  
 Median :65504   Median :100.60   Median :381427   Median :3144   Median :2718  
 Mean   :65317   Mean   :101.68   Mean   :387698   Mean   :3193   Mean   :2607  
 3rd Qu.:68290   3rd Qu.:111.25   3rd Qu.:454086   3rd Qu.:3842   3rd Qu.:3061  
 Max.   :70551   Max.   :116.90   Max.   :554894   Max.   :4806   Max.   :3594  
    POPGT14            YEAR     
 Min.   :107608   Min.   :1947  
 1st Qu.:111788   1st Qu.:1951  
 Median :116804   Median :1954  
 Mean   :117424   Mean   :1954  
 3rd Qu.:122304   3rd Qu.:1958  
 Max.   :130081   Max.   :1962  
```

```
> round(cor(longley),2)
         TOTEMPL INFLAT  GNP UNEMPL ARMYEMPL POPGT14 YEAR
TOTEMPL     1.00   0.97 0.98   0.50     0.46    0.96 0.97
INFLAT      0.97   1.00 0.99   0.62     0.46    0.98 0.99
GNP         0.98   0.99 1.00   0.60     0.45    0.99 1.00
UNEMPL      0.50   0.62 0.60   1.00    -0.18    0.69 0.67
ARMYEMPL    0.46   0.46 0.45  -0.18     1.00    0.36 0.42
POPGT14     0.96   0.98 0.99   0.69     0.36    1.00 0.99
YEAR        0.97   0.99 1.00   0.67     0.42    0.99 1.00
```

According to the summary of the data, we can see that there are no unusual observations because the mean and median of each variable are about the same. From the correlation test, we can see that all variables are highly correlated.

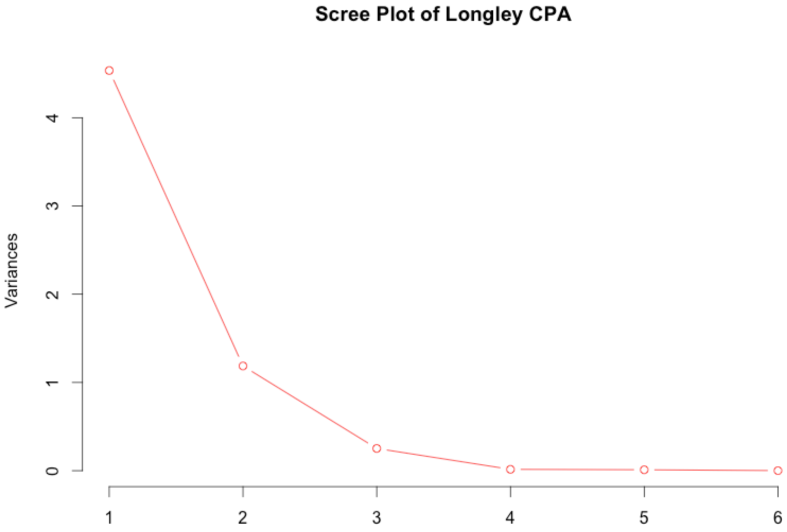
Statement of Problem

We want to find the number principal components that can explain all the data. And we want to find the significant predictors of total employment since all variables are closely correlated.

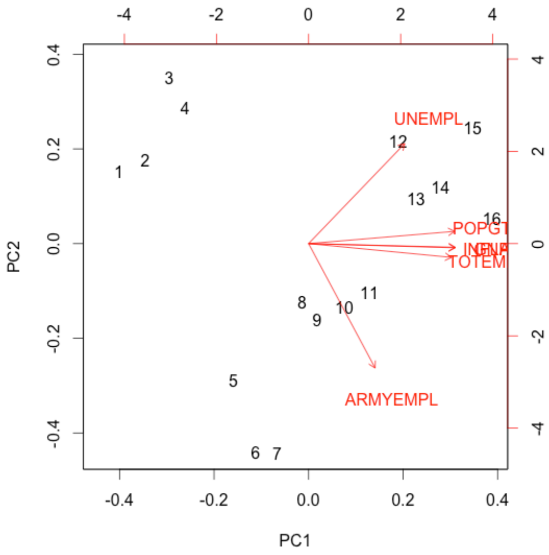
Principal Component Analysis

```
> pca=prcomp(longley[-7], scale. =T)
> summary(pca)
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6
Standard deviation     2.1296 1.0894 0.50190 0.12328 0.10205 0.02656
Proportion of Variance 0.7559 0.1978 0.04198 0.00253 0.00174 0.00012
Cumulative Proportion  0.7559 0.9536 0.99561 0.99815 0.99988 1.00000

screeplot(pca, type="lines",col=3, ,main = "Scree Plot of Longley CPA")
```



From the cumulative proportion of components and scree plot, we can conclude that the first two components are sufficient to explain about 96% of variation in the data.



From the plot PC1 vs PC2, we can see that total employment is related to population over 14, DNP, and inflation.

```
> pca$rotation[,1:2]
               PC1         PC2
TOTEMPL  0.4557966 -0.08589854
INFLAT   0.4669549 -0.02628724
GNP      0.4674899 -0.02306569
UNEMPL   0.3064647  0.62227098
ARMYEMPL 0.2120061 -0.77353962
POPGT14  0.4656056  0.07624745
```

Conclusion I

From the loading of principal component analysis, we can tell that the first component weights TOTEMPL, INFLAT, DNP, and POPGT14 about the same explained 76% of variation in the data. The second principal component mainly measures UNEMPL and ARMEMPL (negatively related to UNEMPL) explained about 20% of variation in the data. Therefore, 2 components explained 96% of variations are enough for all *longley* data.

Variable Selection

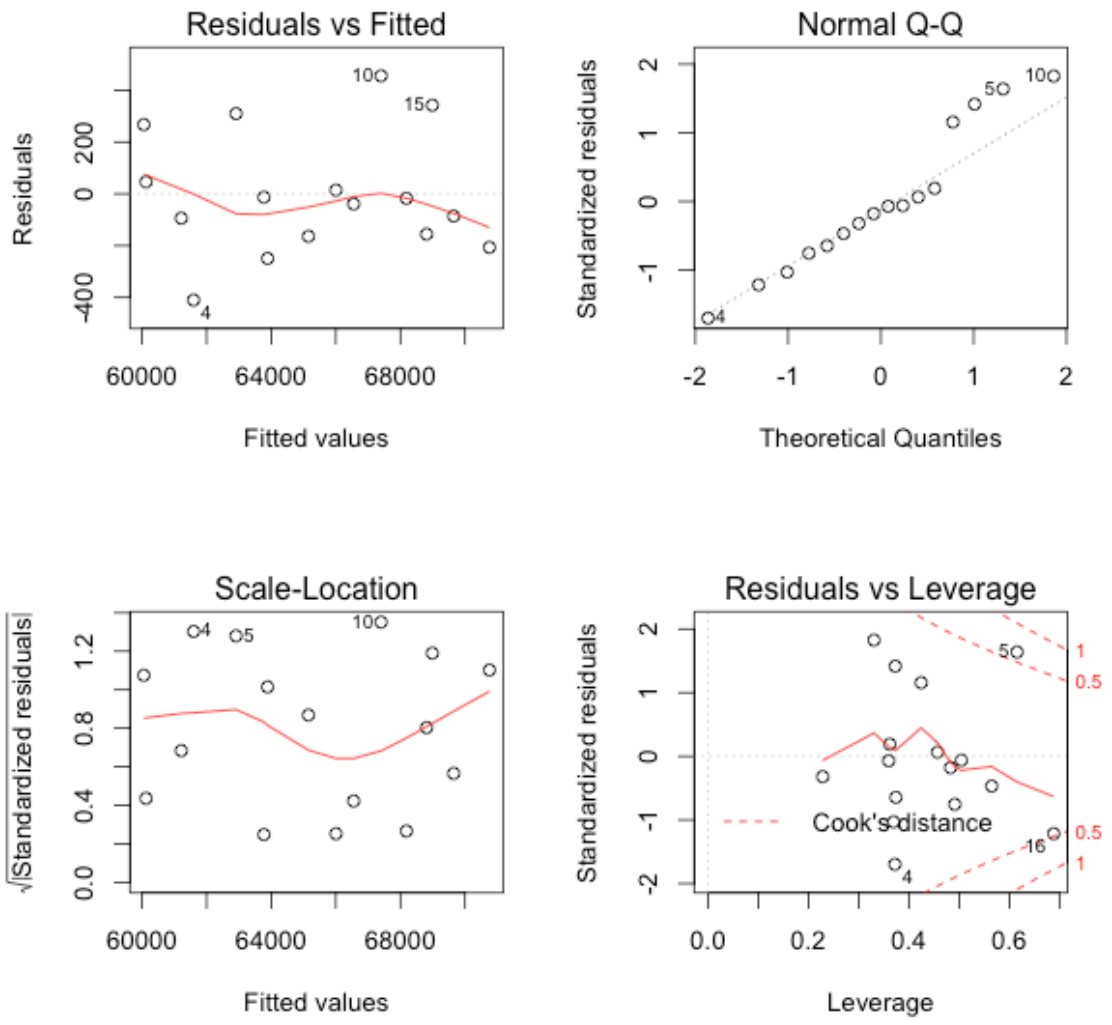
Frist, we fit the data with all the predictors. Then we check the fit of the model by checking for constant errors, normality of errors, and outliners of fitted model.

```
> fit = lm(TOTEMPL ~., longley )
> summary(fit)

Call:
lm(formula = TOTEMPL ~ ., data = longley)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.482e+06  8.904e+05  -3.911 0.003560 **
INFLAT       1.506e+01  8.491e+01   0.177 0.863141
GNP         -3.582e-02  3.349e-02  -1.070 0.312681
UNEMPL      -2.020e+00  4.884e-01  -4.136 0.002535 **
ARMYEMPL    -1.033e+00  2.143e-01  -4.822 0.000944 ***
POPGT14     -5.110e-02  2.261e-01  -0.226 0.826212
YEAR         1.829e+03  4.555e+02   4.016 0.003037 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 304.9 on 9 degrees of freedom
Multiple R-squared:  0.9955,  Adjusted R-squared:  0.9925
```



The model with all predictors (INFLAT, GNP, UNEMPL, ARMYEMPL, POPGT14 and YEAR) has a high $R^{2.}$ From Residual vs. Fitted plot, the errors are constant. From the normal qq plot, the errors are normally distributed. From residuals vs, leverage plot, there are two outliners based on the Cook's distance at significant level 0.05, therefore, we removed observation 5 and 16 and fit a new model.

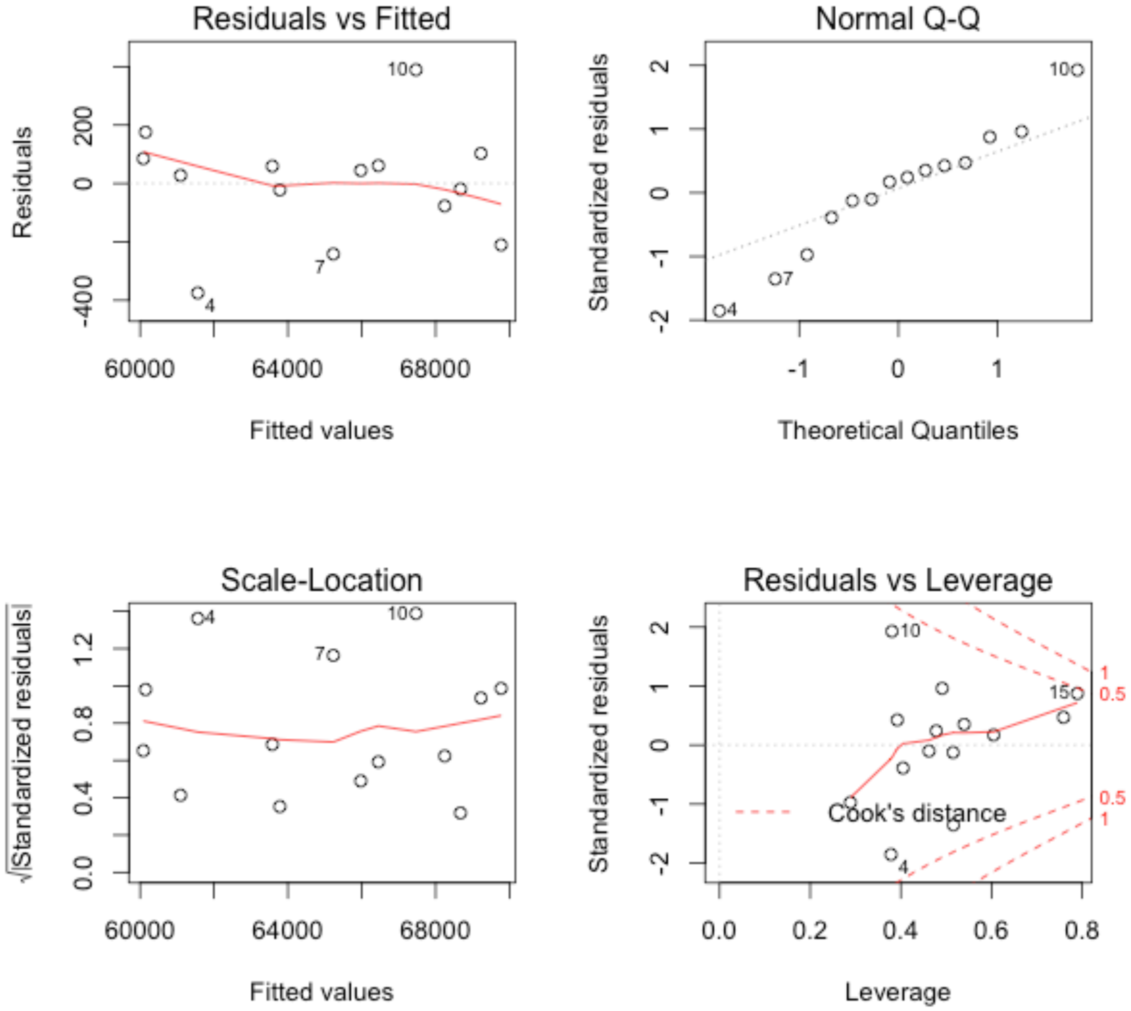
```
> fit = lm(TOTEMPL ~., longley, subset = -c(5,16))
> summary(fit)

Call:
lm(formula = TOTEMPL ~ ., data = longley, subset = -c(5, 16))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -4.485e+06  1.121e+06  -4.000  0.00519 **
INFLAT      -2.920e+00  7.611e+01  -0.038  0.97046
GNP         -7.460e-02  3.799e-02  -1.964  0.09031 .
UNEMPL      -2.618e+00  5.414e-01  -4.836  0.00189 **
ARMYEMPL    -1.179e+00  2.205e-01  -5.348  0.00107 **
POPGT14      2.600e-01  2.306e-01   1.127  0.29670
YEAR         2.333e+03  5.739e+02   4.066  0.00478 **
---

Residual standard error: 256.4 on 7 degrees of freedom
Multiple R-squared:  0.997,   Adjusted R-squared:  0.9944
```

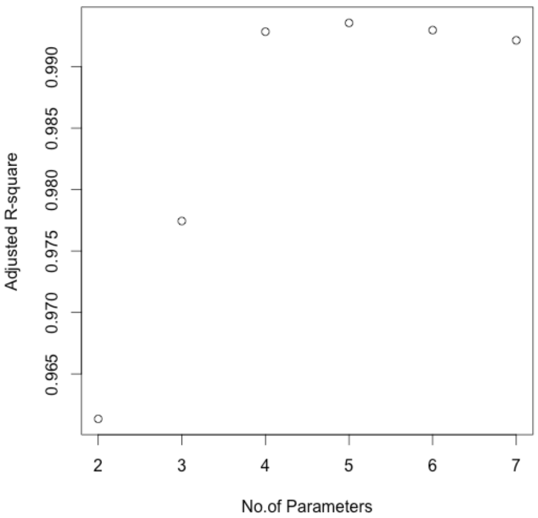


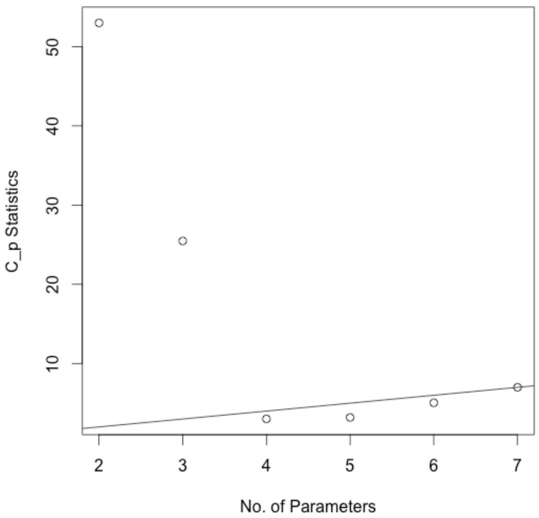
After outliners are removed, the adjusted R^2 increased by a little. From Residual vs. Fitted plot, the errors are constant. From the normal qq plot, the errors are normally distributed. From residuals vs, leverage plot, there are no outliners based on the Cook's distance at significant level 0.05. Then, we use Mallow's C_p Statistics to find significant predictors.

```
> library(leaps)
> all<- regsubsets(TOTEMPL~., data=longley, subset = -c(5,16))
> (rs<-summary(all))
Subset selection object
Call: regsubsets.formula(TOTEMPL ~ ., data = longley)

Selection Algorithm: exhaustive
         INFLAT GNP UNEMPL ARMYEMPL POPGT14 YEAR
1  ( 1 ) " "    "*" " "    " "      " "     " " 
2  ( 1 ) " "    " " "*"    " "      " "     "*" 
3  ( 1 ) " "    " " "*"    "*"      " "     "*" 
4  ( 1 ) " "    "*" "*"    "*"      " "     "*" 
5  ( 1 ) " "    "*" "*"    "*"      "*"     "*" 
6  ( 1 ) "*"    "*" "*"    "*"      "*"     "*" 

> plot(2:7,rs$adjr2, xlab="No.of Parameters", ylab="Adjusted R-square")
> plot(2:7,rs$cp, xlab="No. of Parameters", ylab="C_p Statistics")
> abline(0,1)
```





Based on the Mallow's C_p statistics, 4 parameters are chose with R^2 approximated 1. Therefore, we would include UNEMPL, AMRYEMPL, and YEAR as predictors of TOTEMPL. Then, we check the significant of each variable in the best-fitted model.

```
> fit3 = lm(TOTEMPL ~ UNEMPL + ARMYEMPL + YEAR, longley, subset = -c(5,
16))
> summary(fit3)

Call:
lm(formula = TOTEMPL ~ UNEMPL + ARMYEMPL + YEAR, data = longley,
    subset = -c(5, 16))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.859e+06  6.726e+04 -27.647 8.89e-11 ***
UNEMPL      -1.537e+00  1.519e-01 -10.116 1.43e-06 ***
ARMYEMPL    -8.601e-01  1.706e-01  -5.042 0.000505 ***
YEAR         9.885e+02  3.478e+01  28.417 6.77e-11 ***

Residual standard error: 295.4 on 10 degrees of freedom
Multiple R-squared:  0.9943,  Adjusted R-squared:  0.9926
```

Conclusion II

From the summary of the best-fitted model, we can see that UNEMPL, ARMYEMPL, and YEAR are significant predictors of TOTEMPL from 1947 to 1962 with fitted model R^2 approximate 1.