HW 5

Data *longley*, model with *Employed* as the response and the other variables as predictors. The high Rˆ2 & a majority of predicotr being insignificant could be a potential signal for collinearity.

```
##
## Call:
## lm(formula = Employed ~ GNP.deflator + GNP + Unemployed + Armed.Forces +
##     Population + Year, data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41011 -0.15767 -0.02816  0.10155  0.45539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
## GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
## GNP          -3.582e-02  3.349e-02  -1.070 0.312681
## Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
## Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
## Population   -5.110e-02  2.261e-01  -0.226 0.826212
## Year          1.829e+00  4.555e-01   4.016 0.003037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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
```

1. There are 3 condition number (relative size of eigenvalues) that are greater than the accepted 30 threshold. These results are indicative that some predictors are linear combinations of others, and XˆtX is singular/close to singular. This causes a 'lack of identifiability' in which there is no unique least squares estimate of b, or if there is one than it is imprecise. The standard errors are inflated so that t-tests may fail to reveal significant factors. The fit becomes very sensitive to measurement errors where small changes in y can lead to large changes in b_hat. The solution may require removing some predictors.

```
## [1]    1.00000   17.85504   25.15256   60.78472 1647.47771 5751.21560
```

2. As hinted by the previous results, we do find high correlations with different set of predictors. **GNP.deflator** has a high positive correlation with **GNP**, **Population**, and **Year**. **GNP** has a high positive correlation with **GNP.deflator**,**Population**, and **Year**. **Population** has a high positive correlation with **GNP.deflator**, **GNP**, and **Year**. Finally, **Year** has a high positive correlation with **GNP.deflator**,**GNP**, and **Population**. In short, **GNP.deflator**,**GNP**, **Year**, and **Population** are extremely correlated.

```
##              GNP.deflator  GNP Unemployed Armed.Forces Population Year
## GNP.deflator         1.00 0.99       0.62         0.46       0.98 0.99
## GNP                  0.99 1.00       0.60         0.45       0.99 1.00
## Unemployed           0.62 0.60       1.00        -0.18       0.69 0.67
## Armed.Forces         0.46 0.45      -0.18         1.00       0.36 0.42
## Population           0.98 0.99       0.69         0.36       1.00 0.99
## Year                 0.99 1.00       0.67         0.42       0.99 1.00
```

3. Since we found multiple condition numbers greater than 30, we expect that problems are being caused by more than just one linear combination (more than one set of predictors). The variance inflation factors (VIFs) allows us to quantify the standard error for a particular predictor. It is noticable that the highly correlated predictors mentioned earlier, have the highest VIF (i.e. GNP.deflator, GNP, Population, & Year). For example, we can interpret sqrt(758.98) = 27.5496 as telling us that the standard error for Year is 27.5496 times larger than it would have been without collinearity. We cannot apply this as a correction because we did not actually observe orthogonal data, but it does give us a sense of the size of the effect.

```
## GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
##    135.53244   1788.51348     33.61889      3.58893    399.15102    758.98060
```

4. We have too many variables that are trying to do the same job of explaining the response. We can reduce the collinearity by carefully removing some of the variables. But we should not conclude that the variables we drop have nothing to do with the response. Since GNP.deflator, GNP, Population, & Year are extremely correlated with eachother- —any one of them might do a good job of representing the other. It make sense to use only one of them in our model. I have picked **GNP** for simplicity.

   Comparing this with the original fit, we see that the fit is very similar in terms of R2,but fewer predictors are used. The coefficients are all significant (alpha=10%). The condition numbers are all below the advised 30 threshold. There seem to be no extreme correlations (although Unemployed & GNP is indeed a bit high). Finally, the VIF's seem to be within the range of 1 (orthogonal predictors ==1). Overall, this model is superior to the previous in terms of simplicity & statistical 'soundness.'

```
##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces, data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83085 -0.22306  0.01735  0.10699  1.08090
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.306461   0.716342  74.415  < 2e-16 ***
## GNP           0.040788   0.002207  18.485 3.49e-10 ***
## Unemployed   -0.007968   0.002134  -3.734  0.00285 **
## Armed.Forces -0.004828   0.002552  -1.892  0.08286 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4793 on 12 degrees of freedom
## Multiple R-squared:  0.9851, Adjusted R-squared:  0.9814
## F-statistic: 264.4 on 3 and 12 DF,  p-value: 3.189e-11
```

```
## [1]  1.000000  6.849526 14.436288
```

```
##               GNP Unemployed Armed.Forces
## GNP          1.00       0.60         0.45
## Unemployed   0.60       1.00        -0.18
## Armed.Forces 0.45      -0.18         1.00
```

```
##          GNP   Unemployed Armed.Forces
##     3.140867     2.596610     2.058847
```