Stat 500 - Homework 4 (Solutions)

0. Before we fit a linear model with `Employed` as response and all other variables as regressors, let's take a look at the data.

```
> library(faraway)
> data(longley)
> names(longley)
[1] "GNP.deflator" "GNP"          "Unemployed"   "Armed.Forces"
[5] "Population"   "Year"         "Employed"
```

Inspecting the names of the other variables, we can already conclude that there will be collinearity: It is reasonable to expect that `GNP.deflator` and `Population` are correlated with `GNP`. Moreover, digging a little deeper and inspecting the actual data, we see that the population grows steadily over the years. Therefore, we can expect that variables `Year` and `Population` are (highly) correlated as well.

Those of you with a background in economics might also recognize an endogeneity problem (employment and unemployment are functionally dependent on each other), but that's a separate issue and unrelated to collinearity.

The following statistical analysis confirms our intuition about collinearity:

```
> fit <- lm(Employed~.,data=longley)
> X <- model.matrix(fit)[,-1] # '-1' because we discard the intercept
>
> # Condition number
> e <- eigen(t(X) %*% X)
> round(sqrt(e$val[1]/e$val), 2)[length(e$val)]
[1] 5751.22
>
> # Correlation
> round(cor(X), 2)
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
>
> # Variance inflation factor
> round(vif(X), 2)
GNP.deflator          GNP   Unemployed Armed.Forces   Population
135.53      1788.51        33.62         3.59       399.15
Year
758.98
```

1. The condition number is large (much larger than the suggested critical value of about 30). Thus, the inner product of the design matrix $X'X$ is close to singular. Among others this reduces the accuracy of the estimated regression vector and associated standard errors.

2. As predicted variables `GNP`, `GNP.deflator`, `Population`, and `Year` are highly correlated which explains the high conditioning number.
3. The variance inflation factors of `GNP`, `GNP.deflator`, `Population`, and `Year` are extremely large. For example, $\sqrt{VIF(\text{GNP.deflator})} = \sqrt{135.53} = 11.64$ means that the standard error for the coefficient of `GNP.deflator` 11.64 times as large as it would be if `GNP.deflator` were uncorrelated with the other regressors.
4. Since the variables `GNP`, `GNP.deflator`, `Population`, and `Year` are all highly correlated, they all carry the same (amount of) information. Therefore, we can simply refit our linear model with just one of the four variables and the remaining two variables `Unemployed` and `Armed.Forces`. I decided to keep `GNP.deflator`. For this specific data set, I recommend to not use variable `Year` at all, because years (as numbers) and employment rates are functionally unrelated: years can only increase but employment rates can fluctuate. As we can see below, this solves the multicollinearity problem:

```
> fit2 <- lm(Employed~.,data=longley[,c(1,3,4,7)])
> X2 <- model.matrix(fit2)[,-1]
> e2 <- eigen(t(X2) %*% X2)
> round(sqrt(e2$val[1]/e2$val), 2)[length(e2$val)]
[1] 43.28
> round(cor(X2), 2)
                 GNP.deflator Unemployed Armed.Forces
GNP.deflator             1.00       0.62         0.46
Unemployed               0.62       1.00        -0.18
Armed.Forces             0.46      -0.18         1.00
> round(vif(X2), 2)
GNP.deflator   Unemployed Armed.Forces
3.65         2.96         2.32
```

Comparing the fit of the full and the reduced model we see that the estimates differ significantly and that all regressors included in the reduced model are significant at least at the 10% level.

```
> # Full model with multicollinearity problems
> summary(fit)

Call:
lm(formula = Employed ~ ., 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

> # Reduced model without mulitcollinearity
> summary(fit2)

Call:
lm(formula = Employed ~ ., data = longley[, c(1, 3, 4, 7)])

Residuals:
Min       1Q   Median       3Q      Max
-0.81870 -0.46282  0.07278  0.15816  1.18427

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.393412   1.864285  16.303 1.49e-09 ***
GNP.deflator  0.398076   0.030995  12.843 2.26e-08 ***
Unemployed   -0.010725   0.003220  -3.330   0.0060 **
Armed.Forces -0.008165   0.003829  -2.132   0.0543 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```