Homework 4

Sarah Adilijiang

Problem 1

```
library(faraway)
data(longley)
model = lm(Employed~., longley)
summary(model)
```

```
##
## 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
```

(a) condition numbers

```
X = model.matrix(model)[,-1]
e = eigen(t(X)%*%X)
e$values
```

```
## [1] 6.665299e+07 2.090730e+05 1.053550e+05 1.803976e+04 2.455730e+01
## [6] 2.015117e+00
```

```
sqrt(e$values[1]/e$values)
```

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

Answer:

There is a very large range in the eigenvalues. And there are three large condition numbers, which are greater than 30. This means that the highly collinearity problems are being caused by more than just one linear combination within the predictor matrix $X^T X$.

(b) correlations between predictors

```
round(cor(longley),3)
```

```
##              GNP.deflator   GNP Unemployed Armed.Forces Population  Year
## GNP.deflator        1.000 0.992      0.621        0.465      0.979 0.991
## GNP                 0.992 1.000      0.604        0.446      0.991 0.995
## Unemployed          0.621 0.604      1.000       -0.177      0.687 0.668
## Armed.Forces        0.465 0.446     -0.177        1.000      0.364 0.417
## Population          0.979 0.991      0.687        0.364      1.000 0.994
## Year                0.991 0.995      0.668        0.417      0.994 1.000
## Employed            0.971 0.984      0.502        0.457      0.960 0.971
##              Employed
## GNP.deflator    0.971
## GNP             0.984
## Unemployed      0.502
## Armed.Forces    0.457
## Population      0.960
## Year            0.971
## Employed        1.000
```

Answer:

There are several very large pairwise correlations (close to one) both between four predictors (GNP.deflator, GNP, Population, and Year) and between these predictors and the response, which reveals highly pairwise collinearities. The predictor Unemployed is less but still relatively highly correlated with these four predictors.

This result suggests us that we should only keep one of the four strongly correlated variables - GNP.deflator, GNP, Population, and Year - in the model to avoid the multicollinearity problem.

(c) variance inflation factors

```
vif(model)
```

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

Answer:

There is much variance inflation. In practice it is common to say that VIF greater than 5 is problematic. So in this dataset there is a huge collinearity issue. Each of the predictors - GNP.deflator, GNP, Population and Year - are highly explained by the other predictors. The predictor Unemployed is less but still well explained by the other predictors as well.

Problem 2

```
library(faraway)
data(prostate)
model = lm(lpsa~., prostate)
summary(model)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

(a) condition numbers

```
X = model.matrix(model)[,-1]
e = eigen(t(X)%*%X)
e$values
```

```
## [1] 4.790826e+05 6.190704e+04 2.109042e+02 1.756329e+02 6.479853e+01
## [6] 4.452379e+01 2.023914e+01 8.093145e+00
```

```
sqrt(e$values[1]/e$values)
```

```
## [1]   1.00000   2.78186  47.66094  52.22787  85.98499 103.73114 153.85414
## [8] 243.30248
```

Answer:

There is a large range in the eigenvalues. And there are six large condition numbers, which are greater than 30. This means that the highly collinearity problems are being caused by more than just one linear combination within the predictor matrix $X^T X$.

(b) correlations between predictors

```
round(cor(prostate),3)
```

```
##         lcavol lweight   age   lbph    svi    lcp gleason pgg45  lpsa
## lcavol   1.000   0.194 0.225  0.027  0.539  0.675   0.432 0.434 0.734
## lweight  0.194   1.000 0.308  0.435  0.109  0.100  -0.001 0.051 0.354
## age      0.225   0.308 1.000  0.350  0.118  0.128   0.269 0.276 0.170
## lbph     0.027   0.435 0.350  1.000 -0.086 -0.007   0.078 0.078 0.180
## svi      0.539   0.109 0.118 -0.086  1.000  0.673   0.320 0.458 0.566
## lcp      0.675   0.100 0.128 -0.007  0.673  1.000   0.515 0.632 0.549
## gleason  0.432  -0.001 0.269  0.078  0.320  0.515   1.000 0.752 0.369
## pgg45    0.434   0.051 0.276  0.078  0.458  0.632   0.752 1.000 0.422
## lpsa     0.734   0.354 0.170  0.180  0.566  0.549   0.369 0.422 1.000
```

Answer:

There are several reletively large pairwise correlations between predictors (lcavol, lcp), (svi, lcp), (lcp, pgg45), and (gleason, pgg45), which reveals their pairwise collinearities. The response lpsa is also highly correlated with the predictor lcavol.

It's not clear if we should remove some of the highly correlated variables, but it's worthy trying to remov the predictor lcp, or pgg45, or both of them, then compare the reduced model with the full model to see if the reduced model fits better.

(c) variance inflation factors

```
vif(model)
```

```
##   lcavol  lweight      age     lbph      svi      lcp  gleason    pgg45
## 2.054115 1.363704 1.323599 1.375534 1.956881 3.097954 2.473411 2.974361
```

Answer:

However, there is no significant variance inflation prolem. All the VIF's are smaller than 5.

Problem 3

```
# original data and model
library(faraway)
data(longley)
model = lm(Employed~., data = longley)
summary(model)$coefficients[2,]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.01506187 0.08491493 0.17737603 0.86314083
```

```
# bootstrap data 1000 times
beta_GNPdeflator_hat = NULL
SE_beta_GNPdeflator_hat = NULL

for (i in 1:1000){
    indices = sample(1:16, 16, replace = TRUE)
    boot_data = longley[indices, ]
    boot_model = lm(Employed~., data = boot_data)
    beta_GNPdeflator_hat[i] = summary(boot_model)$coefficients[2,1]
    SE_beta_GNPdeflator_hat[i] = summary(boot_model)$coefficients[2,2]
}
```

```
## Warning in summary.lm(boot_model): essentially perfect fit: summary may be
## unreliable

## Warning in summary.lm(boot_model): essentially perfect fit: summary may be
## unreliable

## Warning in summary.lm(boot_model): essentially perfect fit: summary may be
## unreliable

## Warning in summary.lm(boot_model): essentially perfect fit: summary may be
## unreliable
```

```
# compute the empirical mean of beta_GNPdeflator_hat
mean(beta_GNPdeflator_hat)
```

```
## [1] 0.01938958
```

```r
# compute the empirical standard deviation of beta_GNPdeflator_hat
sd(beta_GNPdeflator_hat)
```

```
## [1] 0.1532288
```

```r
# compute the median of bootstrap estimates SE_beta_GNPdeflator_hat
median(SE_beta_GNPdeflator_hat)
```

```
## [1] 0.07434091
```

Answer:

The empirical mean of $\hat{\beta}_{GNP.deflator}$ from bootstrap samples does not match the original model estimate of $\hat{\beta}_{GNP.deflator}$, which is 0.01506187. So the estimate of $\hat{\beta}_{GNP.deflator}$ is biased here.

The median value of SE($\hat{\beta}_{GNP.deflator}$) from bootstrap samples also does not match the empirical standard deviation of $\hat{\beta}_{GNP.deflator}$. So the usual estimate of the SE for $\hat{\beta_{GNP.deflator}}$ does not estimate the variability appropriately here.

These serious problems with the estimation of β and associated quantities are caused by highly collinearity problems of the data set. Collinearity will lead to imprecise estimates of β, so the the estimate of $\hat{\beta}_{GNP.deflator}$ is biased. And we have $var\hat{\beta}_j = \sigma^2 \frac{1}{1-R_j^2} \frac{1}{S_{x_j x_j}}$, since the variable GNP.deflator is highly correlated with other variables thus being highly explained by other variables, the $R^2_{GNP.deflator}$ is large so the $var\hat{\beta}_{GNP.deflator}$ is inflated.

Problem 4

see next page

STAT 34300 HW4

Problem 4:

(a) $\mathrm{Var}(\hat{y}) = \sigma^2\left(1 + x^T (X^TX)^{-1} x\right)$ $\quad \mathrm{Var}(\hat{y}_{-j}) = \sigma^2\left(1 + x_{-j}^T (X_{-j}^T X_{-j})^{-1} x_{-j}\right)$

(b) $(X^TX)^{-1} = \begin{pmatrix} X_{-j}^T X_{-j} & X_{-j}^T X_j \\ X_j^T X_{-j} & X_j^T X_j \end{pmatrix}^{-1} = \begin{pmatrix} X_{-j}^T X_{-j} & X_{-j}^T X_j \\ (X_{-j}^T X_j)^T & X_j^T X_j \end{pmatrix}^{-1} \geq \begin{pmatrix} (X_{-j}^T X_{-j})^{-1} & 0 \\ 0 & 0 \end{pmatrix}$

x_j is the entry j in vector x

$\therefore\ x^T(X^TX)^{-1}x = (x_{-j}^T\ x_j)(X^TX)^{-1}\begin{pmatrix} x_{-j} \\ x_j \end{pmatrix} \geq (x_{-j}^T\ x_j)\begin{pmatrix} (X_{-j}^T X_{-j})^{-1} & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} x_{-j} \\ x_j \end{pmatrix}$

$= x_{-j}^T (X_{-j}^T X_{-j})^{-1} x_{-j}$

$\Rightarrow\ \sigma^2\left(1 + x^T(X^TX)^{-1}x\right) \geq \sigma^2\left(1 + x_{-j}^T (X_{-j}^T X_{-j})^{-1} x_{-j}\right)$

$\Rightarrow\ \mathrm{Var}(\hat{y}) \geq \mathrm{Var}(\hat{y}_{-j})$

(c) X_j is orthogonal to X_k (for all $k \neq j$) $\Rightarrow$ $X_{-j}^T X_j = 0$ $\quad X_j^T X_{-j} = 0$

and when $x_j = 0$

$x^T(X^TX)^{-1}x = (x_{-j}^T\ x_j)\begin{pmatrix} X_{-j}^T X_{-j} & X_{-j}^T X_j \\ X_j^T X_{-j} & X_j^T X_j \end{pmatrix}^{-1}\begin{pmatrix} x_{-j} \\ x_j \end{pmatrix} = (x_{-j}^T\ 0)\begin{pmatrix} X_{-j}^T X_{-j} & 0 \\ 0 & X_j^T X_j \end{pmatrix}^{-1}\begin{pmatrix} x_{-j} \\ 0 \end{pmatrix}$

$= (x_{-j}^T\ 0)\begin{pmatrix} (X_{-j}^T X_{-j})^{-1} & 0 \\ 0 & (X_j^T X_j)^{-1} \end{pmatrix}\begin{pmatrix} x_{-j} \\ 0 \end{pmatrix} = x_{-j}^T (X_{-j}^T X_{-j})^{-1} x_{-j}$

$\Rightarrow\ \sigma^2\left(1 + x^T(X^TX)^{-1}x\right) = \sigma^2\left(1 + x_{-j}^T (X_{-j}^T X_{-j})^{-1} x_{-j}\right)$

$\Rightarrow\ \mathrm{Var}(\hat{y}) = \mathrm{Var}(\hat{y}_{-j})$

(d) i.) a symmetric $n \times n$ real matrix M is said to be positive-semidefinite if $z^T M z \geq 0$ for every non-zero column vector z of n real numbers.

$\Rightarrow$ if $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ is positive-semidefinite, we have $z^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} z \geq 0$ for all $z \in R^n$ (non-zero column)

let $z = \begin{pmatrix} \sqrt{\varepsilon} z_1 \\ -\varepsilon^{-\frac{1}{2}} z_2 \end{pmatrix}$, where $z_1 \in \mathbb{R}^{n\text{col}(A) = n\text{col}(B^T)}$, $z_2 \in \mathbb{R}^{n\text{col}(B) = n\text{col}(C)}$, $\varepsilon > 0$

$$\Rightarrow z^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} z = (\sqrt{\varepsilon} z_1^T \quad -\varepsilon^{-\frac{1}{2}} z_2^T) \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} \sqrt{\varepsilon} z_1^T \\ -\varepsilon^{-\frac{1}{2}} z_2^T \end{pmatrix} = \varepsilon z_1^T A z_1 - z_2^T B^T z_1 - z_1^T B z_2 + \varepsilon^{-1} z_2^T C z_2$$

and for $\begin{pmatrix} \varepsilon A & -B \\ -B^T & \varepsilon^{-1} C \end{pmatrix}$: $(z_1^T \quad z_2^T) \begin{pmatrix} \varepsilon A & -B \\ -B^T & \varepsilon^{-1} C \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \varepsilon z_1^T A z_1 - z_2^T B^T z_1 - z_1^T B z_2 + \varepsilon^{-1} z_2^T C z_2$

$$= (\sqrt{\varepsilon} z_1^T \quad -\varepsilon^{-\frac{1}{2}} z_2^T) \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \begin{pmatrix} \sqrt{\varepsilon} z_1^T \\ -\varepsilon^{-\frac{1}{2}} z_2^T \end{pmatrix}$$

$\Rightarrow$ if $\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}$ is positive-semidefinite, then $\begin{pmatrix} \varepsilon A & -B \\ -B^T & \varepsilon^{-1} C \end{pmatrix}$ is positive-semidefinite for any $\varepsilon > 0$

ii.) from i), we have $(z_1^T \quad z_2^T) \begin{pmatrix} \varepsilon A & -B \\ -B^T & \varepsilon^{-1} C \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \varepsilon z_1^T A z_1 - z_2^T B^T z_1 - z_1^T B z_2 + \varepsilon^{-1} z_2^T C z_2 \geq 0$

for all $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in R^n$ (non-zero column).

if we want for all $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$: $(z_1^T \quad z_2^T) \begin{pmatrix} \varepsilon A & -B \\ -B^T & cI - C \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \geq 0$

we only need that for all $z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$: $z_2^T (cI - C) z_2 \geq z_1^T (\varepsilon^{-1} C) z_2$

$$\Rightarrow \quad z_2^T \cdot cI \cdot z_2 \geq z_1^T (1 + \varepsilon^{-1}) C z_2$$

$$\Rightarrow \quad \frac{c}{1+\varepsilon^{-1}} z_2^T z_2 \geq z_1^T C z_2$$

So if we let $\frac{c}{1+\varepsilon^{-1}} \geq$ largest eigenvalue of $C = \lambda_{\text{largest}}$

i.e. $c \geq (1+\varepsilon^{-1}) \cdot \lambda_{\text{largest}}$ then $\begin{pmatrix} \varepsilon A & -B \\ -B^T & cI - C \end{pmatrix} \geq 0$

iii.) from ii). we have $\begin{pmatrix} \varepsilon A & -B \\ -B^T & cI - C \end{pmatrix} = \begin{pmatrix} (1+\varepsilon) A & 0 \\ 0 & cI \end{pmatrix} - \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \geq 0$

$$\Rightarrow \begin{pmatrix} (1+\varepsilon) A & 0 \\ 0 & cI \end{pmatrix} \geq \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} \geq \begin{pmatrix} (1+\varepsilon) A & 0 \\ 0 & cI \end{pmatrix}^{-1} = \begin{pmatrix} \frac{1}{1+\varepsilon} A^{-1} & 0 \\ 0 & c^{-1} I \end{pmatrix}$$

$$\because \begin{pmatrix} \frac{1}{1+\varepsilon} A^{-1} & 0 \\ 0 & c^{-1} I \end{pmatrix} \geq \begin{pmatrix} \frac{1}{1+\varepsilon} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} \quad \Rightarrow \begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} \geq \begin{pmatrix} \frac{1}{1+\varepsilon} A^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$