Chapter 12

Example: The **longley** dataset is provide in the base version of R and is a macroeconomic data set which provides a well-known example for a highly collinear regression. This dataset includes 7 economical variables, observed yearly from 1947 to 1962 (n=16). For more information on the variables, enter **?longley** into R.

Let's begin by looking at the header of the dataset and calculating the correlation matrix.

```
data(longley)

# show first few observations
head(longley)

##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
## 1947         83.0 234.289      235.6        159.0    107.608 1947   60.323
## 1948         88.5 259.426      232.5        145.6    108.632 1948   61.122
## 1949         88.2 258.054      368.2        161.6    109.773 1949   60.171
## 1950         89.5 284.599      335.1        165.0    110.929 1950   61.187
## 1951         96.2 328.975      209.9        309.9    112.075 1951   63.221
## 1952         98.1 346.999      193.2        359.4    113.270 1952   63.639

# round the correlations to two decimal places for better viewing
round(cor(longley),2)

##              GNP.deflator  GNP Unemployed Armed.Forces Population Year Employed
## GNP.deflator         1.00 0.99       0.62         0.46       0.98 0.99     0.97
## GNP                  0.99 1.00       0.60         0.45       0.99 1.00     0.98
## Unemployed           0.62 0.60       1.00        -0.18       0.69 0.67     0.50
## Armed.Forces         0.46 0.45      -0.18         1.00       0.36 0.42     0.46
## Population           0.98 0.99       0.69         0.36       1.00 0.99     0.96
## Year                 0.99 1.00       0.67         0.42       0.99 1.00     0.97
## Employed             0.97 0.98       0.50         0.46       0.96 0.97     1.00
```

For this example, we will use **Employed** (number of people employed) as the response variable and the other six variables as predictor variables.

Let's begin by using all six predictor variables and fitting the multiple regression model.

```
model1 = lm(Employed~GNP.deflator+GNP+Unemployed+Armed.Forces+Population+Year,
           data=longley)
summary(model1)

##
## 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
```

As can be seen in the above output, three variables (**GNP.deflator**, **GNP**, and **Population**) are not significant.

Let's take a stepwise approach, remove the variable with the largest p-value that is not statistically significant, and refit the model. In this case, let's remove the predictor variable **GNP.deflator** and refit the multiple regression model.

```r
model2 = lm(Employed~GNP+Unemployed+Armed.Forces+Population+Year,
           data=longley)
summary(model2)

##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Population +
##     Year, data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43015 -0.15399 -0.01832  0.10081  0.44964
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.450e+03  8.282e+02  -4.165 0.001932 **
## GNP          -3.196e-02  2.420e-02  -1.321 0.216073
## Unemployed   -1.972e-02  3.861e-03  -5.108 0.000459 ***
## Armed.Forces -1.020e-02  1.908e-03  -5.345 0.000326 ***
## Population   -7.754e-02  1.616e-01  -0.480 0.641607
## Year          1.814e+00  4.253e-01   4.266 0.001648 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2897 on 10 degrees of freedom
## Multiple R-squared:  0.9955,Adjusted R-squared:  0.9932
## F-statistic: 438.8 on 5 and 10 DF,  p-value: 2.242e-11
```

The **Population** variable is again statistically not significant and has the largest p-value, so let's remove that from the model.

```
model3 = lm(Employed~GNP+Unemployed+Armed.Forces+Year,
           data=longley)
summary(model3)

##
## Call:
## lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,
##     data = longley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42165 -0.12457 -0.02416  0.08369  0.45268
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.599e+03  7.406e+02  -4.859 0.000503 ***
## GNP          -4.019e-02  1.647e-02  -2.440 0.032833 *
## Unemployed   -2.088e-02  2.900e-03  -7.202 1.75e-05 ***
## Armed.Forces -1.015e-02  1.837e-03  -5.522 0.000180 ***
## Year          1.887e+00  3.828e-01   4.931 0.000449 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2794 on 11 degrees of freedom
## Multiple R-squared:  0.9954,Adjusted R-squared:  0.9937
## F-statistic: 589.8 on 4 and 11 DF,  p-value: 9.5e-13
```

All of the predictor variables are now statistically significant at $\alpha = 0.05$.

Note that $R^2 = 0.9954$, so 99.54% of the variability in the number of employed people can be explained by the linear relationship with gross national product (GNP), number of unemployed people, number of people in the armed forces, and year. This is a very large value for R^2 which suggests a very good model fit.

One caution is that our predictor variables are highly correlated (see correlation matrix at the beginning of the example), so this could result in high standard errors. This concern of **multicollinearity** is beyond the scope of this class and is covered in more advanced statistics courses.