Lab 12

David Wiley / Duy Truong

March 26, 2019

Using the Hald cement data in Table B.21 and the tools learned today in class, determine if multi-collinearity is present or not. If it exists, identify the variables that possibly cause collinearity.

```
# DATA

dat=read.csv("/home/david/Documents/2019 Spring/Applied Regression/Labs_HW/Data_Sets2/Appendices/data-ta
y = dat$y
x1 = dat$x_1
x2 = dat$x_2
x3 = dat$x_3
x4 = dat$x_4

fit = lm(y~x1+x2+x3+x4, dat)
summary(fit)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1750 -1.6709  0.2508  1.3783  3.9254
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  62.4054    70.0710   0.891   0.3991
## x1            1.5511     0.7448   2.083   0.0708 .
## x2            0.5102     0.7238   0.705   0.5009
## x3            0.1019     0.7547   0.135   0.8959
## x4           -0.1441     0.7091  -0.203   0.8441
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.446 on 8 degrees of freedom
## Multiple R-squared:  0.9824, Adjusted R-squared:  0.9736
## F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

Using the eigensystem analysis of $X'X$ (denoted $\lambda_1, \lambda_2, ..., \lambda_p$), we measure multicollinearity:

$$k = \frac{\lambda_{max}}{\lambda_{min}}$$

Where: $k < 100$, no serious problem $100 < k < 1000$, moderate to strong multicollinearity $k > 1000$, strong multicollinearity

```
stdx = scale(dat[,3:6])
exx = eigen( t(stdx)%*%stdx)
exx
```

```
## eigen() decomposition
## $values
## [1] 26.82844842 18.91279284  2.23927379  0.01948495
##
## $vectors
##            [,1]       [,2]       [,3]      [,4]
## [1,] -0.4759552  0.5089794  0.6755002 0.2410522
## [2,] -0.5638702 -0.4139315 -0.3144204 0.6417561
## [3,]  0.3940665 -0.6049691  0.6376911 0.2684661
## [4,]  0.5479312  0.4512351 -0.1954210 0.6767340
```

```r
max(exx$values)/min(exx$values)
```

```
## [1] 1376.881
```

Since we know there is multicollinearity, we need to find which regressors are involved. To do that we need to measure each eigenvalue condition index:

$$k = \frac{\lambda_{max}}{\lambda_j}$$

```r
max(exx$values)/exx$values
```

```
## [1]    1.000000    1.418534   11.980870 1376.880621
```

We measure linear dependency by analyzing if the condition index for a regressor is larger than 1000. From the data we can see that the fourth regressor has a condition index of 1376.881. This leads us to believe that the fourth regressor is involved with multicollinearity. The rest of the regressors have indices significantly lower than 1000.