R Notebook

library(readr)  
Sleep = read.csv("https://raw.githubusercontent.com/JA-McLean/STOR455/master/data/SleepStudy.csv  
")  
library(Stat2Data)  
data("CountyHealth") ##from Stat2Data

1. Create a linear model using GPA as the response and AlcoholUse, DepressionStatus, Happiness and Drinks as predictors.

mod1 = lm(GPA~AlcoholUse+Drinks+DepressionStatus+Happiness, data=Sleep)  
summary(mod1)

##   
## Call:  
## lm(formula = GPA ~ AlcoholUse + Drinks + DepressionStatus + Happiness,   
## data = Sleep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.32735 -0.23308 0.03205 0.23060 0.85236   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.442522 0.148267 23.218 < 2e-16 \*\*\*  
## AlcoholUseHeavy 0.227005 0.152703 1.487 0.138   
## AlcoholUseLight 0.114326 0.087170 1.312 0.191   
## AlcoholUseModerate 0.148325 0.097688 1.518 0.130   
## Drinks -0.035619 0.008522 -4.180 4.07e-05 \*\*\*  
## DepressionStatusnormal -0.060470 0.076218 -0.793 0.428   
## DepressionStatussevere -0.194575 0.144806 -1.344 0.180   
## Happiness -0.002487 0.005029 -0.494 0.621   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.391 on 245 degrees of freedom  
## Multiple R-squared: 0.09052, Adjusted R-squared: 0.06454   
## F-statistic: 3.484 on 7 and 245 DF, p-value: 0.001405

1. Compute the VIF values for the model

library(car)

## Loading required package: carData

vif(mod1)

## GVIF Df GVIF^(1/(2\*Df))  
## AlcoholUse 2.029196 3 1.125176  
## Drinks 2.007391 1 1.416824  
## DepressionStatus 1.279483 2 1.063552  
## Happiness 1.255598 1 1.120535

Since the vif values are less than 5, this suggests there is not multicollinearity between the predictors, where multicollinearity is when one or more of the predictors is strongly correlated with some combination of the other predictors in the set.

##Using CountyHealth dataset 3. Construct a linear model using the number of Doctors in a county, MDs, as the response and number of hospitals in the county, Hospitals, and the number of beds in the county, Beds, as predictors

mod2 = lm(MDs~Hospitals+Beds, data=CountyHealth)  
summary(mod2)

##   
## Call:  
## lm(formula = MDs ~ Hospitals + Beds, data = CountyHealth)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1238.57 -291.40 79.93 200.92 1511.16   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -472.0761 135.3677 -3.487 0.00103 \*\*   
## Hospitals 117.3967 55.1110 2.130 0.03810 \*   
## Beds 1.2600 0.1435 8.778 1.07e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 494.8 on 50 degrees of freedom  
## Multiple R-squared: 0.9304, Adjusted R-squared: 0.9276   
## F-statistic: 334.1 on 2 and 50 DF, p-value: < 2.2e-16

1. Create a correlation matrix to look for multicollinearity between the Hospitals and Bed predictors

cor(CountyHealth[,3:4])

## Hospitals Beds  
## Hospitals 1.0000000 0.9094098  
## Beds 0.9094098 1.0000000

The correlation is 0.9 which is close to 1 so they are highly correlated and therefore explain the same kind of variability i.e. show multicollinearity.

1. Compute the VIF values for the model

library(car)  
vif(mod2)

## Hospitals Beds   
## 5.781221 5.781221

Since the vif values are greater than 5, this suggests multicollinearity. So you should consider removing one of the variables from the model.