The executives of a company that manufactures backyard antennae want to predict sales by geographic sales district. They believe that the two most important variables in predicting sales are the number of households and the number of owner-occupied households in each district. To develop and test a model, they randomly selected nine districts. For each district, the number of antennae sold in the previous month, the number of households, and the number of owner-occupied households were recorded. The data are shown below where:

\[
\begin{array}{cccc}
\text{District} & y & x_1 & x_2 \\
1 & 50 & 14 & 11 \\
2 & 73 & 28 & 18 \\
3 & 32 & 10 & 5 \\
4 & 121 & 30 & 20 \\
5 & 156 & 48 & 30 \\
6 & 98 & 30 & 21 \\
7 & 62 & 20 & 15 \\
8 & 51 & 16 & 11 \\
9 & 80 & 25 & 17 \\
\end{array}
\]

Multicollinearity in regression exists when the predictor variables \((x)\) are highly correlated with each other. The result of this is the instability of the regression coefficients and deflated \(t\)-statistics (not significant) while the \(F\)-statistic is significant. The problem of multicollinearity does not lower \(R^2\) since by including unneeded variables in the model cannot reduce \(R^2\) (it can only leave it roughly the same). Multicollinearity can be diagnosed with the variance inflation factor \(VIF\) which is defined as

\[
VIF_j = \frac{1}{1 - R^2_j},
\]

where \(R^2_j\) is the \(R^2\) of the regression of the predictor \(x_j\) on all the other predictors. If \(VIF > 10\) then there is multicollinearity. If multicollinearity is present the simplest solution is to remove from the model predictors that are highly correlated with each other.

#Read the data:
```r
> a <- read.table("multicollinearity.txt", header=TRUE)
```

#Correlation matrix:
```r
> cor(a)
```

#Regression of \(y\) on \(x_1\) and \(x_2\):
```r
> q <- lm(a$y ~ a$x1 + a$x2)
> summary(q)
```

Call:
```r
lm(formula = a$y ~ a$x1 + a$x2)
```

Residuals:
```
               Min             1Q          Median          3Q         Max
-17.8516  -2.2027  -0.9256   2.8780   22.4578
```

Coefficients:
```
                     Estimate Std. Error t value  Pr(>|t|)
(Intercept)       -2.382     10.913   -0.218     0.834
a$x1               2.402      2.221    1.081     0.321
a$x2               1.444      3.525    0.409     0.696
```

Residual standard error: 12.1 on 6 degrees of freedom
Multiple R-squared: 0.9279, Adjusted R-squared: 0.9039
F-statistic: 38.63 on 2 and 6 DF, p-value: 0.0003743