Multiple linear regression model and regression diagnostics

Kalyani, Purvesh

December 6, 2022

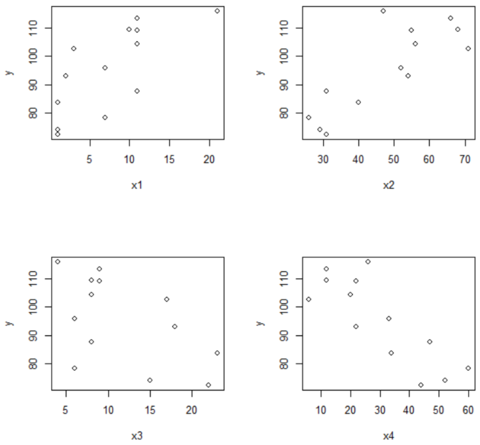
Hald's Cement Data

```
> cement
       y x1 x2 x3 x4
1   78.5  7 26  6 60
2   74.3  1 29 15 52
3  104.3 11 56  8 20
4   87.6 11 31  8 47
5   95.9  7 52  6 33
6  109.2 11 55  9 22
7  102.7  3 71 17  6
8   72.5  1 31 22 44
9   93.1  2 54 18 22
10 115.9 21 47  4 26
11  83.8  1 40 23 34
12 113.3 11 66  9 12
13 109.4 10 68  8 12
```

Description of Variables

- y: heat evolved (calories/gram)
- x1: percentage weight in clinkers of $3CaO.Al_2O_3$.
- x2: percentage weight in clinkers of $3CaO.SiO_2$.
- x3: percentage weight in clinkers of $4CaO.Al_2O_3.Fe_2O_3$.
- x4: percentage weight in clinkers of $2CaO.SiO_2$.

Scatter plots of dependent variables against response variable



Data interpretation

- The y vs x1 plot indicates that the heat evolved increases with increase in percentage weight of $3CaO.Al_2O_3$.
- The y vs x2 plot indicates that the heat evolved increases with increase in percentage weight of $3CaO.SiO_2$.
- The y vs x3 plot shows that there is no specific relationship between the heat evolved and percentage weight of $4CaO.Al_2O_3.Fe_2O_3$.
- The y vs x4 plot indicates that the heat evolved decreases with increase in percentage weight of $3CaO.Al_2O_3$.

Summary of data

```
> summary(cement)
       y               x1              x2              x3       
 Min.   : 72.50   Min.   : 1.000   Min.   :26.00   Min.   : 4.00  
 1st Qu.: 83.80   1st Qu.: 2.000   1st Qu.:31.00   1st Qu.: 8.00  
 Median : 95.90   Median : 7.000   Median :52.00   Median : 9.00  
 Mean   : 95.42   Mean   : 7.462   Mean   :48.15   Mean   :11.77  
 3rd Qu.:109.20   3rd Qu.:11.000   3rd Qu.:56.00   3rd Qu.:17.00  
 Max.   :115.90   Max.   :21.000   Max.   :71.00   Max.   :23.00  
       x4    
 Min.   : 6  
 1st Qu.:20  
 Median :26  
 Mean   :30  
 3rd Qu.:44  
 Max.   :60  
```

Building a model

```
> ml = lm(y ~., data = cement)
> summary(ml)

Call:
lm(formula = y ~ ., data = cement)

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
```

Data interpretation

- The p value of model $= 4.756e^{-07}$ indicates that at least one of the predictor variables is significantly related to the response variable.
- The R-square is 0.9824, meaning that approximately 98% of the variability of y is accounted for by the variables in the model.
- The adjusted R-square shows after taking the account of number of predictors in the model R-square is 0.9736.

Coefficients

- $y = 62.4054 + 1.5511 * x1 + 0.5102 * x2 + 0.1019 * x3 - 0.1441 * x4$
- The p value of t statistic for x3 and x4, $= 0.8959, 0.8441$, is very high. Thus, x3 and x4 are not significantly associated with y.
- Since the p value for x3 is highest, we remove x3 from the model.

```
> anova(m1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1         1 1450.08 1450.08 242.3679 2.888e-07 ***
x2         1 1207.78 1207.78 201.8705 5.863e-07 ***
x3         1    9.79    9.79   1.6370    0.2366    
x4         1    0.25    0.25   0.0413    0.8441    
Residuals  8   47.86    5.98                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- ANOVA table indicates that x1 and x2 are highly significant for the model.

Model 2 excluding x3

```
> m2 = lm(y ~ x1+x2+x4, data = cement)
> summary(m2)

Call:
lm(formula = y ~ x1 + x2 + x4, data = cement)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0919 -1.8016  0.2562  1.2818  3.8982 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
x1            1.4519     0.1170  12.410 5.78e-07 ***
x2            0.4161     0.1856   2.242 0.051687 .  
x4           -0.2365     0.1733  -1.365 0.205395    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,	Adjusted R-squared:  0.9764 
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

Interpretations

- The p value of model $= 3.323e^{-08}$ has decreased indicating better fit.
- The R-square is 0.9823 and adjusted R-square 0.9764 has increased.
- $y = 71.6483 + 1.4519 * x1 + 0.4161 * x2 - 0.2365 * x4$
- The p value of t statistic for x4, $= 0.2054$, is high. So, it is possible to remove x4 from the model.

```
> anova(m2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1         1 1450.08 1450.08 272.0439 4.934e-08 ***
x2         1 1207.78 1207.78 226.5879 1.094e-07 ***
x4         1    9.93    9.93   1.8633    0.2054    
Residuals  9   47.97    5.33                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- ANOVA table indicates that x1 and x2 are highly significant for the model.

Model 3 excluding x3 and x4

```
> m3 = lm(y~ x1 + x2, data = cement)
> summary(m3)

Call:
lm(formula = y ~ x1 + x2, data = cement)

Residuals:
   Min     1Q Median     3Q    Max
-2.893 -1.574 -1.302  1.363  4.048

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.57735    2.28617   23.00 5.46e-10 ***
x1           1.46831    0.12130   12.11 2.69e-07 ***
x2           0.66225    0.04585   14.44 5.03e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.406 on 10 degrees of freedom
Multiple R-squared:  0.9787,    Adjusted R-squared:  0.9744
F-statistic: 229.5 on 2 and 10 DF,  p-value: 4.407e-09
```

Interpretations

- The p value of model $= 4.407e^{-09}$ has decreased indicating better fit.
- The R-square is 0.9787 and adjusted R-square is 0.9744.
- $y = 52.57735 + 1.4683 * x1 + 0.66225 * x2$

```
> anova(m3)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x1         1 1450.1 1450.08  250.43 2.088e-08 ***
x2         1 1207.8 1207.78  208.58 5.029e-08 ***
Residuals 10   57.9    5.79
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

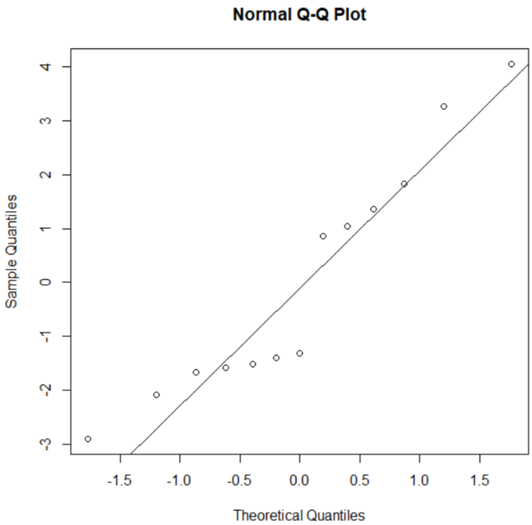
- ANOVA table indicates that x1 and x2 are highly significant for the model.

Regression Diagnostics

We will check the following assumptions on residuals.

- **Normality**: the errors should be normally distributed.
- **Homoscedasticity** (Homogeneity of variance) : The error variance should be constant
- **Linearity:** the relationships between the predictors and the outcome variable should be linear
- **Multicollinearity:** predictors that are highly collinear, i.e., linearly related, can cause problems in estimating the regression coefficients.
- **Independence:** The errors associated with one observation are not correlated with the errors of any other observation
- **Influence:** individual observations that exert undue influence on the coefficients

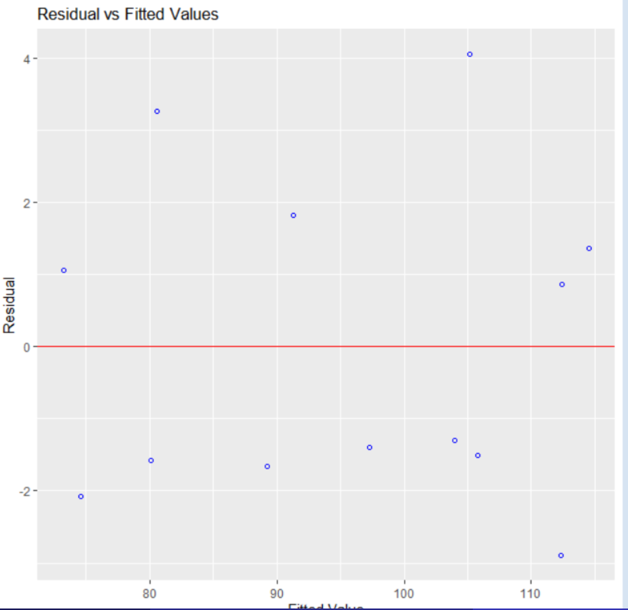
Normality of Residuals



Normality tests

```
> ols_test_normality(m3)
-----------------------------------------------
       Test             Statistic       pvalue
-----------------------------------------------
Shapiro-Wilk              0.9053          0.1580
Kolmogorov-Smirnov        0.2618          0.2825
Cramer-von Mises          1.0053          0.0018
Anderson-Darling          0.5972          0.0953
-----------------------------------------------
```

Homoscedasticity (Homogeneity of variance)



Interpretation

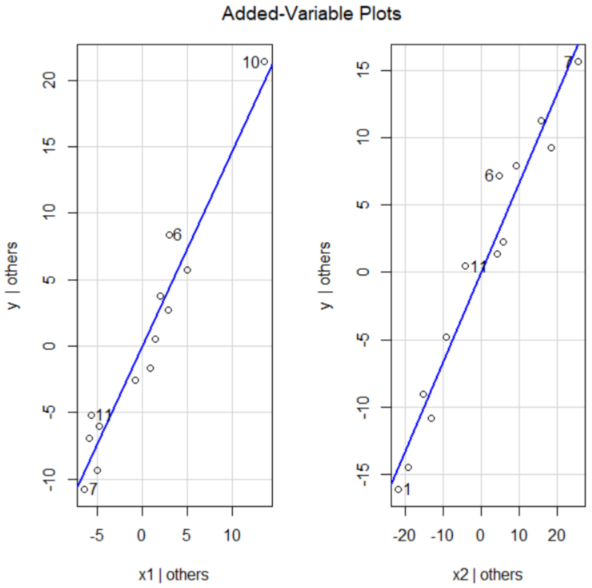
- The residuals can be contained in a horizontal band.
- This indicated that the variance of residuals is approximately constant.
- The fitted line almost follows a straight line, indicating linearity in the model.

Multicollinearity

```
> require(car)
> car::vif(m3)
      x1       x2 
1.055129 1.055129 
```

- The VIF (variance inflation factor) measures the degree of multicollinearity.
- Both x1 and x2 have low VIF values, indicating no multicollinearity.

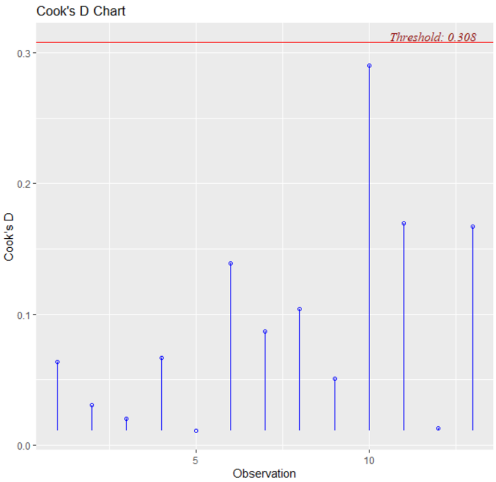
Added Variable Plot



Interpretation

- The added variable plot is scatter plot of residuals of a model by excluding one variable from the full model against residuals of a model that uses the excluded variable as dependent variable predicted by other variables.
- The slope of the simple regression between those residuals will be the same as coefficient of the excluded variable.
- The non zero slope of both plots implies that variables x1 and x2 are relevant to the model.

Cook's Distance for checking influence



- All the Cook's D values are within the threshold limit indicating low possibility of outliers in the model.

Conclusion

- We fitted model $y = 52.57735 + 1.4683 * x1 + 0.66225 * x2$ for Hald's cement data and verified the assumptions on residuals.

Thank You