Predictive Analysis PS 3

Problem 5

Atrijo Roy

2026-02-19

Problem to demonstrate the utility of non-linear regression over linear regression

Get the fgl data set from “MASS” library.

```
#Loading the data
rm(list=ls())
library(MASS)

## Warning: package 'MASS' was built under R version 4.5.2

attach(fgl)
head(fgl)

##      RI    Na   Mg   Al    Si    K   Ca Ba   Fe type
## 1  3.01 13.64 4.49 1.10 71.78 0.06 8.75  0 0.00 WinF
## 2 -0.39 13.89 3.60 1.36 72.73 0.48 7.83  0 0.00 WinF
## 3 -1.82 13.53 3.55 1.54 72.99 0.39 7.78  0 0.00 WinF
## 4 -0.34 13.21 3.69 1.29 72.61 0.57 8.22  0 0.00 WinF
## 5 -0.58 13.27 3.62 1.24 73.08 0.55 8.07  0 0.00 WinF
## 6 -2.04 12.79 3.61 1.62 72.97 0.64 8.07  0 0.26 WinF
```

(a) Considering the refractive index (RI) of “Vehicle Window glass” as the variable of interest and assuming linearity of regression, run multiple linear regression of RI on different metallic oxides. From the p value, report which metallic oxide best explains the refractive index.

```
#Separating the continuous variables
df=fgl[fgl$type=="Veh",]
df$type = NULL
head(df)

##        RI    Na   Mg   Al    Si    K   Ca Ba   Fe
## 147 -0.31 13.65 3.66 1.11 72.77 0.11 8.60  0 0.00
## 148 -1.90 13.33 3.53 1.34 72.67 0.56 8.33  0 0.00
## 149 -1.30 13.24 3.57 1.38 72.70 0.56 8.44  0 0.10
## 150 -1.57 12.16 3.52 1.35 72.89 0.57 8.53  0 0.00
## 151 -1.35 13.14 3.45 1.76 72.48 0.60 8.38  0 0.17
## 152  3.27 14.32 3.90 0.83 71.50 0.00 9.49  0 0.00

#(a)
fit1=lm(RI~.,data=df)
summary(fit1)
```

```
##
## Call:
## lm(formula = RI ~ ., data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29194 -0.08582  0.00072  0.10740  0.33524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 131.4641    47.2669   2.781  0.02388 *
## Na           -0.4333     0.3509  -1.235  0.25190
## Mg           -0.2866     1.0075  -0.285  0.78325
## Al           -0.8909     0.5550  -1.605  0.14713
## Si           -1.8824     0.4993  -3.770  0.00547 **
## K            -2.4232     0.9725  -2.492  0.03743 *
## Ca            1.5326     0.5818   2.634  0.02998 *
## Ba            0.3517     2.6904   0.131  0.89922
## Fe            3.8931     0.9581   4.063  0.00362 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2621 on 8 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9813
## F-statistic: 105.9 on 8 and 8 DF,  p-value: 2.622e-07
```

Fitted Model:

$$\widehat{RI} = 131.4641 - 0.4333\,Na - 0.2866\,Mg - 0.8909\,Al - 1.8824\,Si - 2.4232\,K + 1.5326\,Ca \\ + 0.3517\,Ba + 3.8931\,Fe$$

with R squared = 0.9906, so the fit is very good.

From the p values we can see Fe is the most significant predictor in the multiple linear regression of RI on all the continuous predictors as the p-value corresponding to Fe = 0.00362 is the least among the other p-values.

(b) Run a simple linear regression of RI on the best predictor chosen in (a).

```
fit2=lm(RI~Fe,data=df)
summary(fit2)

##
## Call:
## lm(formula = RI ~ Fe, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2324 -1.0693 -0.2715  0.2907  3.7707
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5007     0.4861  -1.030   0.3193
## Fe            8.1362     4.0780   1.995   0.0645 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.759 on 15 degrees of freedom
## Multiple R-squared:  0.2097, Adjusted R-squared:  0.157
## F-statistic: 3.981 on 1 and 15 DF,  p-value: 0.06452
```

Fitted Model:

$$\widehat{RI} = -0.5007 + 8.1362\, Fe$$

with multiple R square = 0.2097 so the fit is not good.

(c) Can you further improve the regression of the refractive index of “Vehicle Window glass” on the predictor chosen by you in part (a)? Give the new fitted model and compare its performance with the model in (b).

```
#Quadratic model
fit3=lm(RI~Fe+I(Fe^2),data=df)
summary(fit3)
```

```
##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6215 -1.1715 -0.1345  0.5985  3.5485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2785     0.4712  -0.591    0.564
## Fe          -12.1810    12.0408  -1.012    0.329
## I(Fe^2)      65.9600    37.0798   1.779    0.097 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.645 on 14 degrees of freedom
## Multiple R-squared:  0.3554, Adjusted R-squared:  0.2633
## F-statistic:  3.86 on 2 and 14 DF,  p-value: 0.04623
```

Fitted Model:

$$\widehat{RI} = -0.2785 - 12.1810\, Fe + 65.9600\, Fe^2$$

with R squared = 0.3554, the fit is not good but shows clear improvement over linear regression.

```
#Cubic Model
fit4=lm(RI~Fe+I(Fe^2)+I(Fe^3),data=df)
summary(fit4)
```

```
##
## Call:
## lm(formula = RI ~ Fe + I(Fe^2) + I(Fe^3), data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6306 -1.1806 -0.0695  0.5621  3.5394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.2694     0.4921  -0.548    0.593
## Fe          -16.7947    32.2946  -0.520    0.612
## I(Fe^2)     107.1214   268.4871   0.399    0.696
## I(Fe^3)     -79.0070   510.0359  -0.155    0.879
##
## Residual standard error: 1.705 on 13 degrees of freedom
## Multiple R-squared:  0.3566, Adjusted R-squared:  0.2081
## F-statistic: 2.402 on 3 and 13 DF,  p-value: 0.1146
```

Fitted Model:

$$\widehat{RI} = -0.2694 - 16.7947\,Fe + 107.1214\,Fe^2 - 79.0070\,Fe^3$$

with R square = 0.3566

Conclusion:

Quadratic is giving substantial improvement over linear regression but Cubic is slight improvement over quadratic so we choose quadratic regression as improvement over liner regression model.