Practical Exercise 5 | Statistics for CSAI II

Burcu_Ibicioglu, u986202

The goals of this exercise are to (a) to use R to run multiple linear regression models that include polynomials, b) running mixed models, and c) growth curve models.

Tasks indicate things that you need to complete in R/R Studio.

Task 1. Load the winequality-red.csv data file.

```
data<-read.csv('/Users/burcuibicioglu/Downloads/Practical Exercise 5-2/Practical Exercise 5/Wine.csv', s
```

Task 2. Inspect the data by looking at the first few entries and the last few entries in the dataset as well as the variable types. In particular, we are interested in predicting the "quality" of the red wine, by knowing the "total.sulfur.dioxide" content of the wine.

```
head(data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.4             0.70        0.00            1.9     0.076
## 2           7.8             0.88        0.00            2.6     0.098
## 3           7.8             0.76        0.04            2.3     0.092
## 4          11.2             0.28        0.56            1.9     0.075
## 5           7.4             0.70        0.00            1.9     0.076
## 6           7.4             0.66        0.00            1.8     0.075
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  11                   34  0.9978 3.51      0.56     9.4
## 2                  25                   67  0.9968 3.20      0.68     9.8
## 3                  15                   54  0.9970 3.26      0.65     9.8
## 4                  17                   60  0.9980 3.16      0.58     9.8
## 5                  11                   34  0.9978 3.51      0.56     9.4
## 6                  13                   40  0.9978 3.51      0.56     9.4
##   quality
## 1       5
## 2       5
## 3       5
## 4       6
## 5       5
## 6       5
```

```
tail(data)
```

```
##      fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1594           6.8            0.620        0.08            1.9     0.068
## 1595           6.2            0.600        0.08            2.0     0.090
## 1596           5.9            0.550        0.10            2.2     0.062
## 1597           6.3            0.510        0.13            2.3     0.076
```

```
## 1598           5.9            0.645        0.12            2.0     0.075
## 1599           6.0            0.310        0.47            3.6     0.067
##      free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1594                  28                   38 0.99651 3.42      0.82     9.5
## 1595                  32                   44 0.99490 3.45      0.58    10.5
## 1596                  39                   51 0.99512 3.52      0.76    11.2
## 1597                  29                   40 0.99574 3.42      0.75    11.0
## 1598                  32                   44 0.99547 3.57      0.71    10.2
## 1599                  18                   42 0.99549 3.39      0.66    11.0
##      quality
## 1594       6
## 1595       5
## 1596       6
## 1597       6
## 1598       5
## 1599       6
```

```r
summary(data$total.sulfur.dioxide)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   22.00   38.00   46.47   62.00  289.00
```

```r
summary(data$quality)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.636   6.000   8.000
```

a. Generate descriptive statistics. Evaluate these descriptives and print them here.

```r
install.packages("psych")
```

```
## Error in contrib.url(repos, "source"): trying to use CRAN without setting a mirror
```

```r
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```r
describe.by(data)
```

```
## Warning: describe.by is deprecated.  Please use the describeBy function

## Warning in describeBy(x = x, group = group, mat = mat, type = type, ...): no
## grouping variable requested

##                      vars    n  mean    sd median trimmed   mad  min    max
## fixed.acidity           1 1599  8.32  1.74   7.90    8.15  1.48 4.60  15.90
## volatile.acidity        2 1599  0.53  0.18   0.52    0.52  0.18 0.12   1.58
## citric.acid             3 1599  0.27  0.19   0.26    0.26  0.25 0.00   1.00
## residual.sugar          4 1599  2.54  1.41   2.20    2.26  0.44 0.90  15.50
## chlorides               5 1599  0.09  0.05   0.08    0.08  0.01 0.01   0.61
## free.sulfur.dioxide     6 1599 15.87 10.46  14.00   14.58 10.38 1.00  72.00
## total.sulfur.dioxide    7 1599 46.47 32.90  38.00   41.84 26.69 6.00 289.00
## density                 8 1599  1.00  0.00   1.00    1.00  0.00 0.99   1.00
## pH                      9 1599  3.31  0.15   3.31    3.31  0.15 2.74   4.01
## sulphates              10 1599  0.66  0.17   0.62    0.64  0.12 0.33   2.00
## alcohol                11 1599 10.42  1.07  10.20   10.31  1.04 8.40  14.90
## quality                12 1599  5.64  0.81   6.00    5.59  1.48 3.00   8.00
##                       range skew kurtosis   se
## fixed.acidity         11.30 0.98     1.12 0.04
```

```
## volatile.acidity       1.46 0.67     1.21 0.00
## citric.acid            1.00 0.32    -0.79 0.00
## residual.sugar        14.60 4.53    28.49 0.04
## chlorides              0.60 5.67    41.53 0.00
## free.sulfur.dioxide   71.00 1.25     2.01 0.26
## total.sulfur.dioxide 283.00 1.51     3.79 0.82
## density                0.01 0.07     0.92 0.00
## pH                     1.27 0.19     0.80 0.00
## sulphates              1.67 2.42    11.66 0.00
## alcohol                6.50 0.86     0.19 0.03
## quality                5.00 0.22     0.29 0.02
```

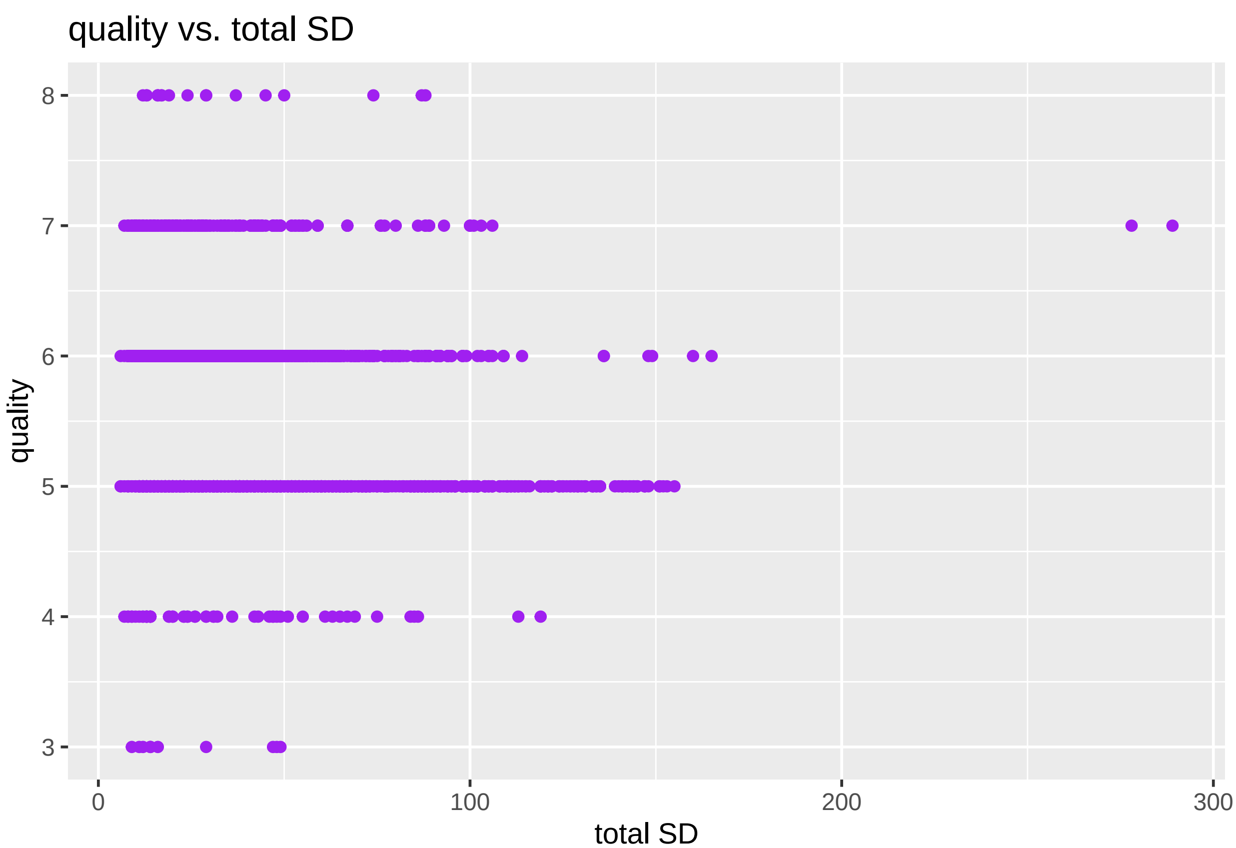
b. Make a scatter plot of the relationship between "quality" and "total.sulfur.dioxide". Does it look like the relationship is best fit by a straight line or perhaps something curvilinear?

```r
library(ggplot2)
```

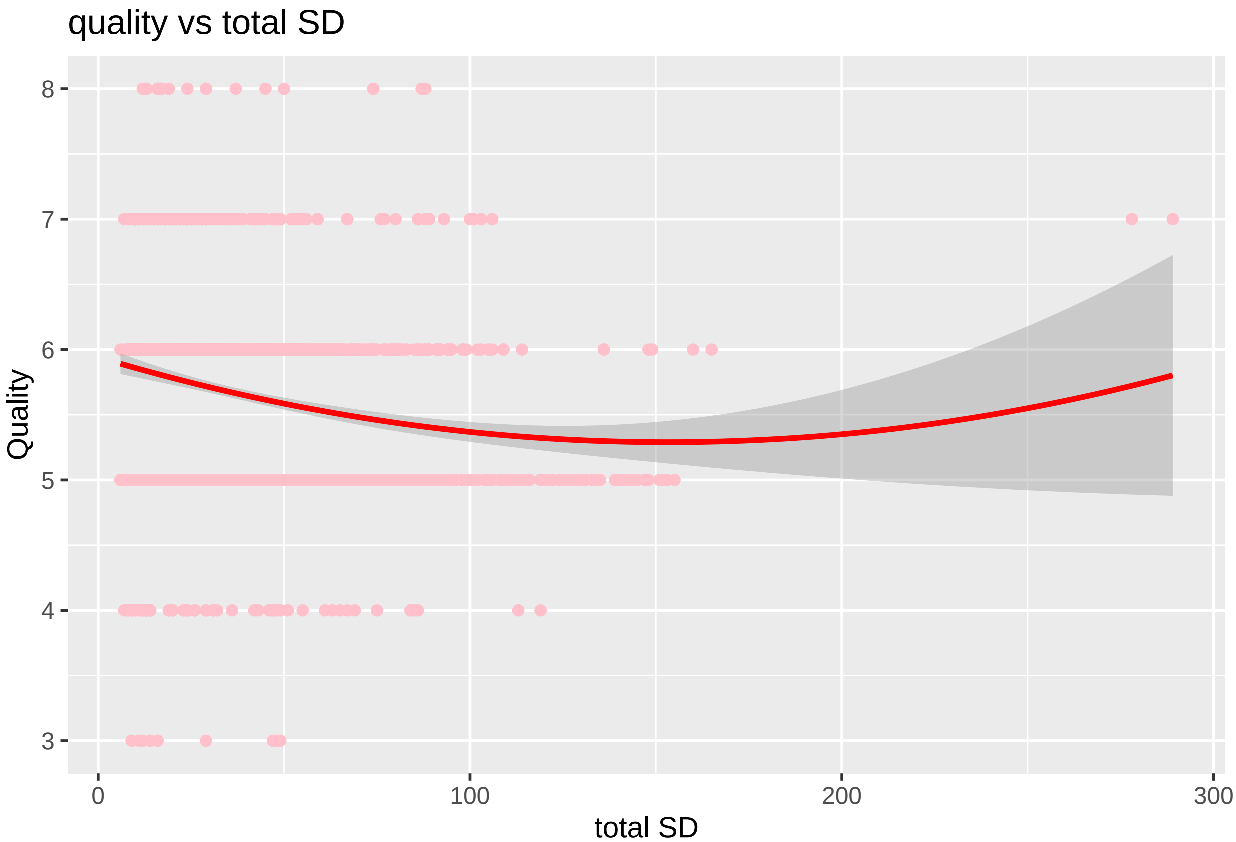
```
##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```r
ggplot(data, aes(x = total.sulfur.dioxide, y = quality)) + geom_point(color = "purple") + ggtitle("quali
```



```r
ggplot(data, aes(x = total.sulfur.dioxide, y = quality)) + geom_point(color = "pink") + geom_smooth(meth
```




```r
##BETTER FIT BY A CURVILINEAR##
```

Task 3. Run a series of polynomial multiple regression models with "quality" as your outcome that includes "total.sulfur.dioxide" as a predictor. Start with a linear model, then add a quadratic term, then run another model that includes a cubic term. Compare the results of the models.

```r
model <- lm(quality ~ total.sulfur.dioxide, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = quality ~ total.sulfur.dioxide, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8063 -0.6336  0.2164  0.3800  2.5527
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.8471792  0.0343670 170.140  < 2e-16 ***
## total.sulfur.dioxide -0.0045442  0.0006037  -7.527 8.62e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7939 on 1597 degrees of freedom
## Multiple R-squared:  0.03426,    Adjusted R-squared:  0.03366
```

```
## F-statistic: 56.66 on 1 and 1597 DF,  p-value: 8.622e-14
```

```r
qmodel <- lm(quality ~ total.sulfur.dioxide + I(total.sulfur.dioxide^2), data = data)
summary(qmodel)
```

```
##
## Call:
## lm(formula = quality ~ total.sulfur.dioxide + I(total.sulfur.dioxide^2),
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8670 -0.6028  0.1723  0.4146  2.5923
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.941e+00  4.773e-02 124.480  < 2e-16 ***
## total.sulfur.dioxide      -8.508e-03  1.521e-03  -5.592 2.64e-08 ***
## I(total.sulfur.dioxide^2)  2.777e-05  9.789e-06   2.837  0.00461 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7921 on 1596 degrees of freedom
## Multiple R-squared:  0.03911,    Adjusted R-squared:  0.0379
## F-statistic: 32.48 on 2 and 1596 DF,  p-value: 1.495e-14
```

```r
cmodel <- lm(quality ~ total.sulfur.dioxide + I(total.sulfur.dioxide^2) + I(total.sulfur.dioxide^3), da
summary(cmodel)
```

```
##
## Call:
## lm(formula = quality ~ total.sulfur.dioxide + I(total.sulfur.dioxide^2) +
##     I(total.sulfur.dioxide^3), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7637 -0.6670  0.2371  0.3459  2.5971
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                5.749e+00  6.722e-02  85.523  < 2e-16 ***
## total.sulfur.dioxide       2.454e-03  3.107e-03   0.790  0.42980
## I(total.sulfur.dioxide^2) -1.087e-04  3.515e-05  -3.092  0.00202 **
## I(total.sulfur.dioxide^3)  4.099e-07  1.014e-07   4.040 5.59e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7883 on 1595 degrees of freedom
## Multiple R-squared:  0.04884,    Adjusted R-squared:  0.04705
## F-statistic:  27.3 on 3 and 1595 DF,  p-value: < 2.2e-16
```

a. Report the results here in APA format. Be sure to include the adjusted R2 value, the b estimates, and the p-values. What can you conclude from your results and which model best characterizes this relationship?

###The results show that the cubic model provided the best fit for the data, with an adjusted

R^2=0.04705.However, the model had a non-significant p-value for the linear term (p=0.4298), which can mean that it may not be statistically meaningful. The quadratic model also showed a significant relationship (p=0.00461), with an adjusted R^2=0.0379. The linear model explained juts 3.36% of the variance, suggesting limited prediction for total sulfur dioxide. In conclusion, while the cubic model provided a better fit in terms of adjusted R^2, the quadratic model is more reliable because it is statistically significant.###