Assginment-7 Multiple Regression

Pradeep Paladugula

6/23/2020

```
Q1. Why are we concerned with multicollinearity?

Solution: Firstly we will understand what is collinearity mean. Two variables
are said perfectly collinear if there is an exact linear relationship between
them.
   Now Multicollinearity means two or more variables in a multiple regression
model are highly linearly related. The key role of regression analysis is to
isolate the relationship between        the independent variable and
dependent variable. We can change the values of the independent variable and
not on others. When independent variables are correlated, It indicates that
the change in one variables are associated with the changes in anothe
variable. The stronger the correlation, the more difficult is to change one
variable without changing other.

    we are concerned with Multicollinearity because these causes:
    1. MC reduces the precision of the estimate coefficiets. It also wekens
the p-value to identify independent variables that are statistically
significant.
    2. The cofficients become very sensitive even for small changes in the
model/data.
    3. It gets difficult to find out the actual effect of each variable.

    using variance inflation factor(VIF) identifies correlation between
independent variables and the strength of that correlation. It is very
simplest test to assess multicolnearity.




Q2. How might we sift through predictors to find the best set of predictors
for our model?

Solution: Predictor variables are also know as independent variables or input
variables.
    1. if we are identifying only one predictor variable for the model, we
can predict the the best independent variable using correlation plot. (NOTE:
(y=mx+b) which of the variables                        have an impact on the
'x' independent variable that correlates with 'y' tehe most.)
    2. If we have to identify multiple predictor variables from the model,
multiple regression analysis is most appropriate to find the most fit
predictors.
```

a) the predictor variable with the largest absolute value for the standardized coefficient.

b) the predictor variable that is associated with the greatest increase in R-squared.

c) Low p-values don’t necessarily identify predictor variables that are practically important.

Q3. If we run an ANOVA(model1, model2) and the p-value is greater than .05 what does this mean?

Solution: If p value is greater than 0.05, then the modal non significant and we should reject the modal from considerations.

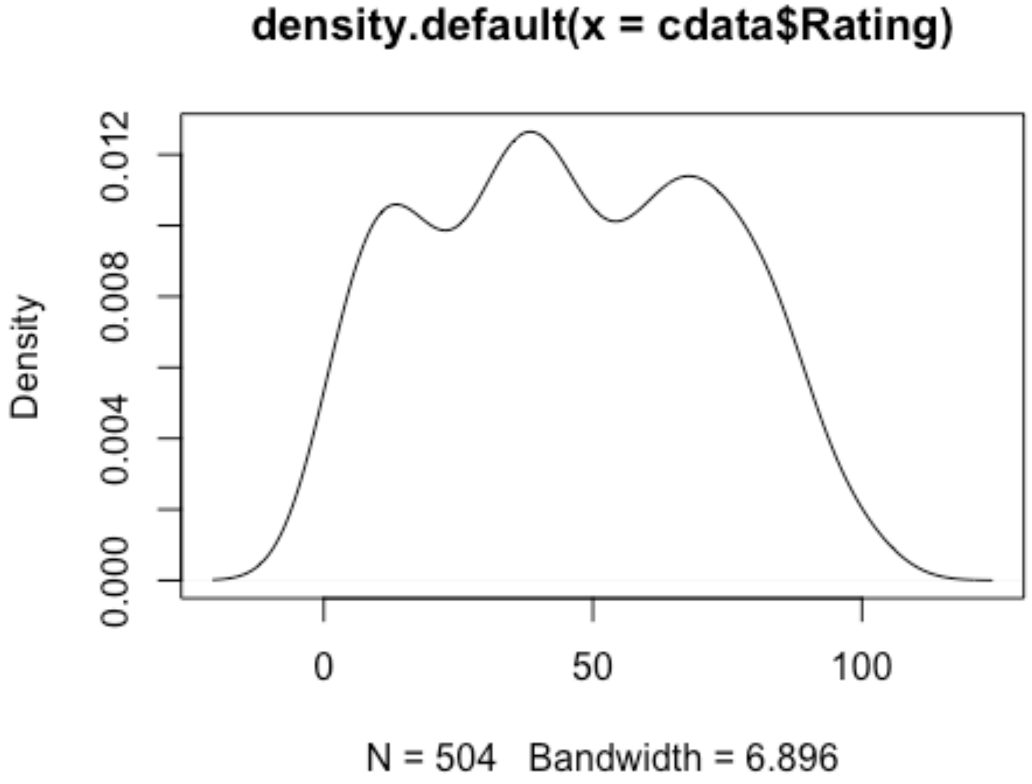
Q4. Use the in-class practice data to write a summary. (Just submit Summary, no code needed for submission) Use "RegressionExample.xlsx" file in Module 8.

Solution: A linear regression model was conducted to predict ad's rating increase, based on the ad's and ages of two different sex groups acting in the ad's. All the regression assumptions were met, and no further adjustments made. A significant regression equation was found (F(2,501) = 1332, p < .001), with an R-square of .84, which suggests that 84% of the variance of ad's rating increased can be explained by the two predictors. Both ad (t = 50.659, p < .001. b = 29.33) and Age (t = -7.82, p < .001, b = -0.023) were statistically significant. The result suggested that ad's predicts that for every number of ad, there is a 29.33% increase in ad's rating. Besides, Age also predicts that for every certain age groups there is only a 0.23% increase in ad's rating. And the negative sign of estiamte value indicates that as the increase in the age groups thier is a decrease in the ad's rating.

Q5. Research Question: Determine whether how many times the ad is watched and if age can significantly predict rating.

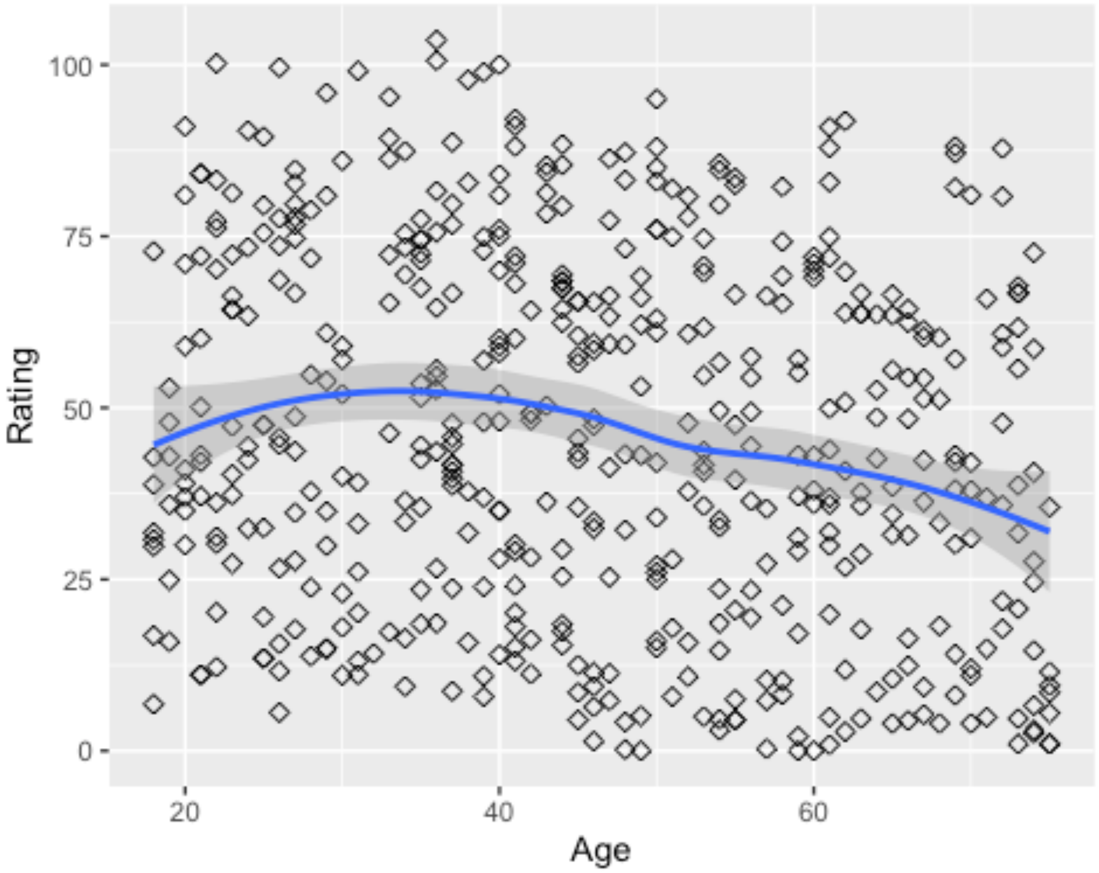
solution: From my prediction: From ANOVA method between model2 and model3. The result shows the DF of 1 (indticating an additional parameter) and very small p-value(<0.001). this means that adding the age IV to the model2 did lead to significantly improved fit over the model2. I conclude that the ad IV variable was watched twice in this analysis.

```
library(xlsx)
library(ggplot2)
library(car)

## Loading required package: carData

library(readxl)
cdata <- read_excel('RegressionExample.xlsx')
cdata

## # A tibble: 504 x 4
##        Ad   Age Sex    Rating
##    <dbl> <dbl> <chr>    <dbl>
##   1     0    55 male      4.5
##   2     0    56 female   19.4
##   3     1    56 male     36.4
##   4     1    54 female   49.6
##   5     2    36 male     75.6
##   6     2    36 female   81.6
##   7     0    75 male      5.5
##   8     0    42 female   28.2
##   9     1    61 male     35.9
##  10     1    41 female   60.1
## # … with 494 more rows

cor(cdata$Rating, cdata$Ad, use = "complete.obs", method = "pearson")

## [1] 0.9068833

cor.test(cdata$Rating, cdata$Ad, method = "pearson")

##
##  Pearson's product-moment correlation
##
## data:  cdata$Rating and cdata$Ad
## t = 48.22, df = 502, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8900406 0.9212536
## sample estimates:
##       cor
## 0.9068833

plot(density(cdata$Rating))
```

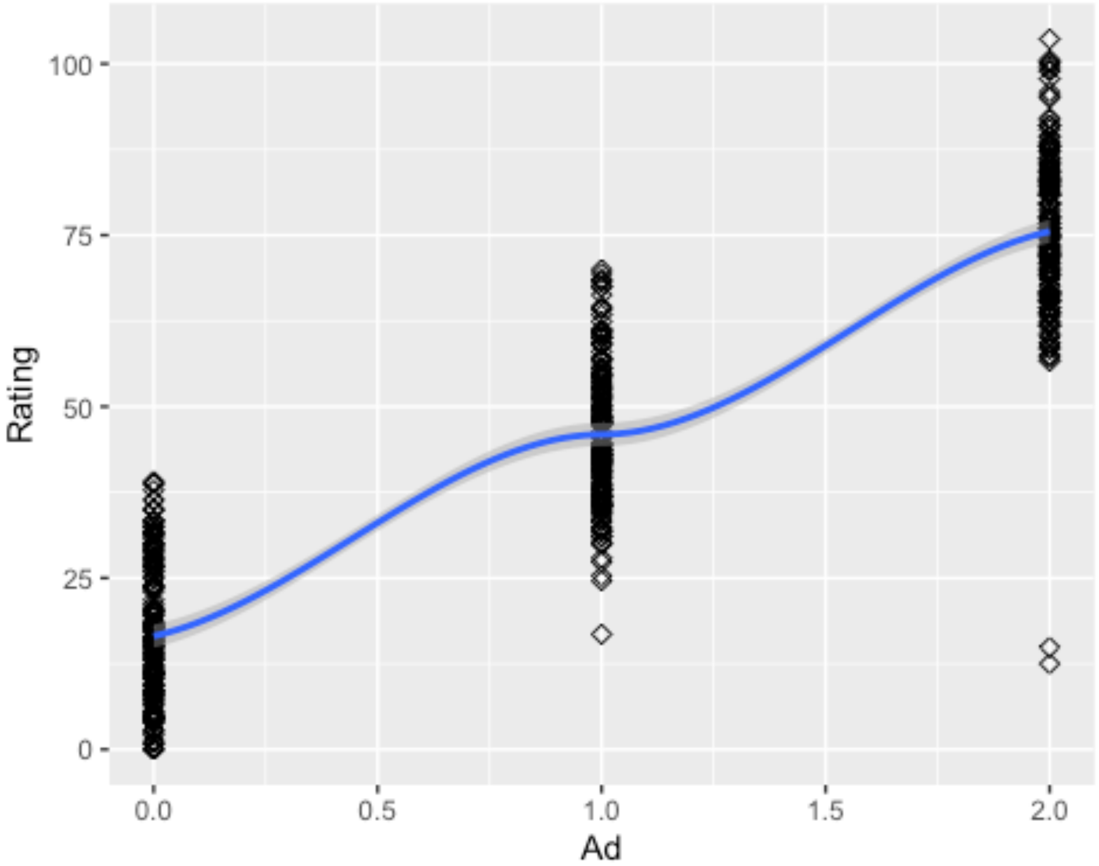


```r
ggplot(cdata, aes(x=Age, y=Rating)) +
  geom_point(size=2, shape=23)+
  geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



```
ggplot(cdata, aes(x=Ad, y=Rating)) +
  geom_point(size=2, shape=23)+
  geom_smooth(method = 'loess')

## `geom_smooth()` using formula 'y ~ x'

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : pseudoinverse used at -0.01

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : neighborhood radius 2.01

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : reciprocal condition number 1.4469e-15

## Warning in simpleLoess(y, x, w, span, degree = degree, parametric =
## parametric, : There are other near singularities as well. 4.0401

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : pseudoinverse used
at
## -0.01
## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : neighborhood
radius 2.01

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : reciprocal
condition
## number 1.4469e-15

## Warning in predLoess(object$y, object$x, newx = if
## (is.null(newdata)) object$x else if (is.data.frame(newdata))
## as.matrix(model.frame(delete.response(terms(object)), : There are other
near
## singularities as well. 4.0401
```



```
model <- lm(Rating ~ Ad*Age, cdata)
model

##
## Call:
## lm(formula = Rating ~ Ad * Age, data = cdata)
##
```

```
## Coefficients:
## (Intercept)           Ad          Age       Ad:Age  
##   27.455351    29.032870    -0.234501     0.006388
```

```
summary(model)
```

```
## 
## Call:
## lm(formula = Rating ~ Ad * Age, data = cdata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -63.043  -7.471  -1.572   7.612  26.061 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 27.455351   2.221037  12.362  < 2e-16 ***
## Ad          29.032870   1.770343  16.400  < 2e-16 ***
## Age         -0.234501   0.044843  -5.229  2.5e-07 ***
## Ad:Age       0.006388   0.036399   0.175    0.861    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 10.61 on 500 degrees of freedom
## Multiple R-squared:  0.8417, Adjusted R-squared:  0.8408 
## F-statistic: 886.5 on 3 and 500 DF,  p-value: < 2.2e-16
```

```
BIC(model)
```

```
## [1] 3838.34
```

```
vif(model)
```

```
##        Ad       Age    Ad:Age 
##  9.349700  2.350816 10.420625
```

```
model2 <- lm(Rating ~ Ad, cdata)
model2
```

```
## 
## Call:
## lm(formula = Rating ~ Ad, data = cdata)
## 
## Coefficients:
## (Intercept)           Ad  
##       16.49        29.51
```

```
summary(model2)
```

```
## 
## Call:
## lm(formula = Rating ~ Ad, data = cdata)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.013  -8.322  -0.951   7.921  28.087
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  16.4878     0.7901   20.87   <2e-16 ***
## Ad           29.5128     0.6120   48.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.22 on 502 degrees of freedom
## Multiple R-squared:  0.8224, Adjusted R-squared:  0.8221
## F-statistic:  2325 on 1 and 502 DF,  p-value: < 2.2e-16
```

```
BIC(model2)
```

```
## [1] 3883.9
```

```
model3 <- lm(Rating ~ Ad + Age, cdata)
model3
```

```
##
## Call:
## lm(formula = Rating ~ Ad + Age, data = cdata)
##
## Coefficients:
## (Intercept)           Ad          Age
##     27.1779      29.3264      -0.2285
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = Rating ~ Ad + Age, data = cdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -63.046  -7.393  -1.536   7.728  25.997
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.17791    1.55844  17.439  < 2e-16 ***
## Ad          29.32643    0.57890  50.659  < 2e-16 ***
## Age         -0.22854    0.02924  -7.815 3.26e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.6 on 501 degrees of freedom
## Multiple R-squared:  0.8417, Adjusted R-squared:  0.8411
## F-statistic:  1332 on 2 and 501 DF,  p-value: < 2.2e-16
```

```
BIC(model3)
```

```
## [1] 3832.148
```

```
vif(model3)
```

```
##     Ad    Age
## 1.0017 1.0017
```

```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.95251, p-value = 1.187e-11
```

```
anova(model2, model3)
```

```
## Analysis of Variance Table
##
## Model 1: Rating ~ Ad
## Model 2: Rating ~ Ad + Age
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    502 63184
## 2    501 56319  1    6865.5 61.075 3.258e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```