R Notebook

# Question 1

## Importing necessary library

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.4 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.4.0 ✓ forcats 0.5.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(scatterplot3d)  
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

## Importing data

First, the data set is imported into the notebook.

data <- read\_csv("Advertising.csv")

## Warning: Missing column names filled in: 'X1' [1]

##   
## ── Column specification ────────────────────────────────────────────────────────  
## cols(  
## X1 = col\_double(),  
## TV = col\_double(),  
## radio = col\_double(),  
## newspaper = col\_double(),  
## Sales = col\_double()  
## )

attach(data)

## Data inspection

Then we have a quick inspect in the data set.

str(data)

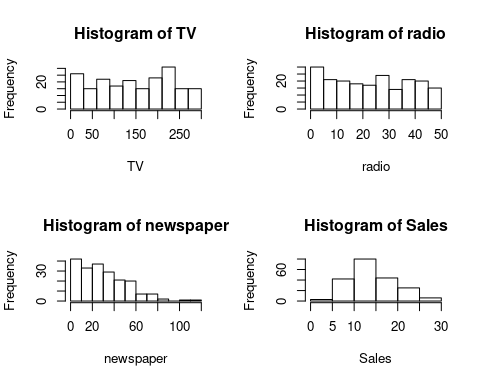
## tibble [200 × 5] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ X1 : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...  
## $ TV : num [1:200] 230.1 44.5 17.2 151.5 180.8 ...  
## $ radio : num [1:200] 37.8 39.3 45.9 41.3 10.8 48.9 32.8 19.6 2.1 2.6 ...  
## $ newspaper: num [1:200] 69.2 45.1 69.3 58.5 58.4 75 23.5 11.6 1 21.2 ...  
## $ Sales : num [1:200] 22.1 10.4 9.3 18.5 12.9 7.2 11.8 13.2 4.8 10.6 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. X1 = col\_double(),  
## .. TV = col\_double(),  
## .. radio = col\_double(),  
## .. newspaper = col\_double(),  
## .. Sales = col\_double()  
## .. )

summary(data)

## X1 TV radio newspaper   
## Min. : 1.00 Min. : 0.70 Min. : 0.000 Min. : 0.30   
## 1st Qu.: 50.75 1st Qu.: 74.38 1st Qu.: 9.975 1st Qu.: 12.75   
## Median :100.50 Median :149.75 Median :22.900 Median : 25.75   
## Mean :100.50 Mean :147.04 Mean :23.264 Mean : 30.55   
## 3rd Qu.:150.25 3rd Qu.:218.82 3rd Qu.:36.525 3rd Qu.: 45.10   
## Max. :200.00 Max. :296.40 Max. :49.600 Max. :114.00   
## Sales   
## Min. : 1.60   
## 1st Qu.:10.38   
## Median :12.90   
## Mean :14.02   
## 3rd Qu.:17.40   
## Max. :27.00

Plot a histogram of Sales, TV, radio and newspaper individually

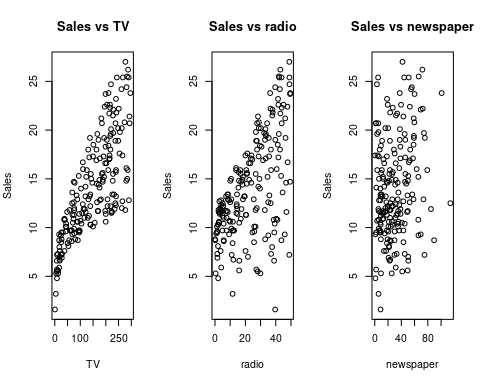
par(mfrow = c(2,2))  
hist(TV,breaks = seq(0,300,30))  
hist(radio)  
hist(newspaper)  
hist(Sales)



The spread of data in TV seems to be even. However, in radio and newspaper, there are higher frequency of lower value investment with newspaper having a more obvious trend. From the histogram, we observed that Sales are roughly normally distributed with slight skewed to the right.

Next, we plot scatter plot of Sales vs TV, radio, and newspaper individually.

par(mfrow = c(1,3))  
plot(TV, Sales, main = "Sales vs TV")  
plot(radio, Sales, main = "Sales vs radio")  
plot(newspaper, Sales, main = "Sales vs newspaper")



## Data pre-processing

Now we randomly shuffle the data.

set.seed(100) # To produce reproducible result  
  
data <- data %>%   
 select(Sales,TV,radio,newspaper) %>%   
 mutate(rand = runif(dim(data)[1]),) %>%   
 arrange(rand)  
  
head(data)

## # A tibble: 6 x 5  
## Sales TV radio newspaper rand  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 5.3 13.1 0.4 25.6 0.0115  
## 2 11.6 48.3 47 8.5 0.0162  
## 3 7.3 11.7 36.9 45.2 0.0190  
## 4 10.9 38 40.3 11.9 0.0196  
## 5 8 17.9 37.6 21.6 0.0267  
## 6 10.6 87.2 11.8 25.9 0.0285

From the scatter plot, we observe that Sales is a concave function of TV. Hence, we will try to fit the data with concave function such as . The relationship between Sales and Radio is roughly linear observing from the graph while there seems to not have any correlation between Sales and Newspaper. We will compute a correlation matrix to find out.

cor(select(data,-rand))

## Sales TV radio newspaper  
## Sales 1.0000000 0.78222442 0.57622257 0.22829903  
## TV 0.7822244 1.00000000 0.05480866 0.05664787  
## radio 0.5762226 0.05480866 1.00000000 0.35410375  
## newspaper 0.2282990 0.05664787 0.35410375 1.00000000

From the correlation matrix, we can see that Sales is highly correlated to TV and Radio with and respectively, while Sales is weakly correlated with Newspaper with . Do note that the predictors Radio and Newspaper are correlated to each other with . Hence, we expect only one of either will make a good predictor.

## Repeated K-fold Cross Validation

### All predictors

Now we will use repeated K-fold cross validation (K = 10) and fit with a linear regression

# First we set training control to repeated K-fold cross validation   
# with K = 10, and repeats = 3  
train.control <- trainControl(method = "repeatedcv",   
 number = 10, repeats = 3)  
  
# Then we train the linear model  
model1 <- train(Sales ~ I(sqrt(TV)) + radio + newspaper, data = data, method = "lm",  
 trControl = train.control)  
  
# Summary of the result  
summary(model1)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.3147 -0.8536 0.0294 0.8485 3.4205   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.6092798 0.3362200 -4.786 3.34e-06 \*\*\*  
## `I(sqrt(TV))` 0.9749476 0.0240384 40.558 < 2e-16 \*\*\*  
## radio 0.1947679 0.0071549 27.222 < 2e-16 \*\*\*  
## newspaper -0.0005253 0.0048803 -0.108 0.914   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.401 on 196 degrees of freedom  
## Multiple R-squared: 0.929, Adjusted R-squared: 0.9279   
## F-statistic: 854.3 on 3 and 196 DF, p-value: < 2.2e-16

print(model1)

## Linear Regression   
##   
## 200 samples  
## 3 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 179, 180, 180, 180, 180, 181, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.400601 0.9330734 1.09849  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

From the statistics we have, and (), this is considered a good model.

However, when we look closely into the t-values of the predictors, we noticed that both sqrt(TV) and radio are significant with while newspaper is not significant with .

The coefficients of the predictors told us that for every 1 additional unit of sqrt(TV) invested, 0.975 unit of Sales will be generated; for every 1 additional unit of radio invested, 0.195 unit of Sales will be generated; for every 1 additional unit of newspaper invested, -0.0005 unit of Sales will be generated, which is a loss.

Now, we will use Variance Inflation Factor(VIF) to check for multicollinearity.

vif(model1$finalModel)

## `I(sqrt(TV))` radio newspaper   
## 1.002183 1.143614 1.144868

From the VIF generated, the , and shows that there is some collinearity between the predictors radio and newspaper.

### Without newspaper

Next we try to train the model without the predictor newspaper.

# Model without newspaper predictor  
model2 <- train(Sales ~ I(sqrt(TV)) + radio, data = data, method = "lm",  
 trControl = train.control)  
  
# Summary of the result  
summary(model2)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.2997 -0.8514 0.0371 0.8599 3.4128   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.617931 0.325651 -4.968 1.46e-06 \*\*\*  
## `I(sqrt(TV))` 0.974854 0.023962 40.683 < 2e-16 \*\*\*  
## radio 0.194496 0.006677 29.131 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.398 on 197 degrees of freedom  
## Multiple R-squared: 0.929, Adjusted R-squared: 0.9282   
## F-statistic: 1288 on 2 and 197 DF, p-value: < 2.2e-16

print(model2)

## Linear Regression   
##   
## 200 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 180, 180, 180, 180, 180, 180, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.403392 0.9357967 1.092344  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

We can see the new model without the predictor newspaper has a higher and , which indicates that this is a slightly better model compared to the model with newspaper. We will do an ANOVA test to confirm.

anova(model1$finalModel, model2$finalModel)

## Analysis of Variance Table  
##   
## Model 1: .outcome ~ `I(sqrt(TV))` + radio + newspaper  
## Model 2: .outcome ~ `I(sqrt(TV))` + radio  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 196 384.86   
## 2 197 384.88 -1 -0.022748 0.0116 0.9144

However, the ANOVA test has a , which means that we cannot reject the null hypothesis that model1 and model2 fits the data equally well. There is no sufficient evidence that the model without the newspaper feature is significantly better.

Of course, we will check again the VIF of the predictors for any multicollinearity.

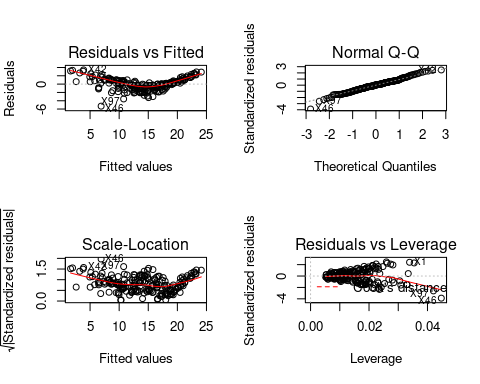
vif(model2$finalModel)

## `I(sqrt(TV))` radio   
## 1.000868 1.000868

We can see both VIF values are approximate equal to 1, which indicates that there are no multicollinearity between the predictors.

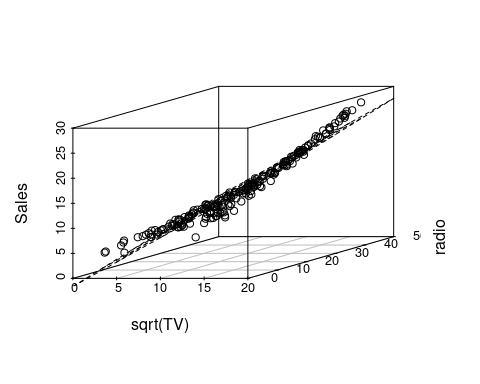
Next up, we will plot the residual graphs to inspect how are the residuals distributed.

par(mfrow = c(2,2))  
plot(model2$finalModel)



We can see that the residuals are not randomly distributed across the 0 abline. This indicates that there are some bias going on with current model. We now try out the 3D scatter plot of the raw and predicted data to inspect.

s3d <- scatterplot3d(x = sqrt(TV), y = radio, z = Sales, angle = 30)  
  
# adding in the prediction plane  
s3d$plane3d(model2$finalModel)



We can see that at the extreme value(pure input) of either radio or sqrt(TV), the model underestimate the data, while when there is a mixed input from both predictors, the model overestimate the data. It clearly shows that there are some synergy or interaction between the 2 predictors term. Hence, we will fit again a model with interaction term between the 2 predictors.

### Adding interaction variables

# Model with interaction term  
model3 <- train(Sales ~ I(sqrt(TV)) \* radio, data = data, method = "lm",  
 trControl = train.control)  
  
# Summary of the result  
summary(model3)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.0562 -0.2757 -0.0121 0.2758 1.2421   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.4444112 0.1793714 24.778 < 2e-16 \*\*\*  
## `I(sqrt(TV))` 0.4383960 0.0150223 29.183 < 2e-16 \*\*\*  
## radio -0.0500957 0.0062645 -7.997 1.09e-13 \*\*\*  
## `I(sqrt(TV)):radio` 0.0215106 0.0005179 41.538 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4476 on 196 degrees of freedom  
## Multiple R-squared: 0.9928, Adjusted R-squared: 0.9926   
## F-statistic: 8949 on 3 and 196 DF, p-value: < 2.2e-16

print(model3)

## Linear Regression   
##   
## 200 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 180, 180, 181, 180, 180, 180, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.4489912 0.9932889 0.3435566  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

We can see that the new model has a and which is way higher than the previous 2 models. We will do ANOVA test to compare this model with model2.

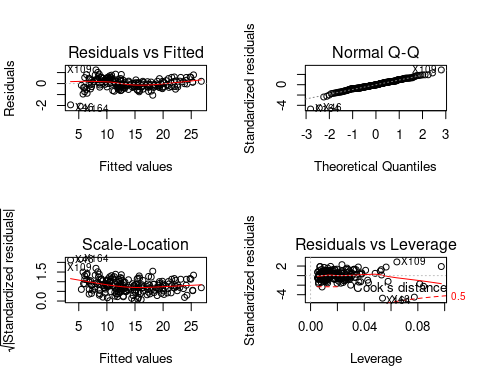
# Comparing with model2  
anova(model2$finalModel, model3$finalModel)

## Analysis of Variance Table  
##   
## Model 1: .outcome ~ `I(sqrt(TV))` + radio  
## Model 2: .outcome ~ `I(sqrt(TV))` + radio + `I(sqrt(TV)):radio`  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 197 384.88   
## 2 196 39.26 1 345.62 1725.4 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

We can see that the of the ANOVA test rejects the null hypothesis that model2 and model3 fits equally well. Hence, there is evidence that the model with interaction term fits significantly better than then the model without the interaction term.

We can also plot the residual plot to see how the residuals are distributed.

par(mfrow = c(2,2))  
plot(model3$finalModel)



Now we can see that the residuals are roughly randomly distributed across the 0 abline.

## Principal Component Analysis

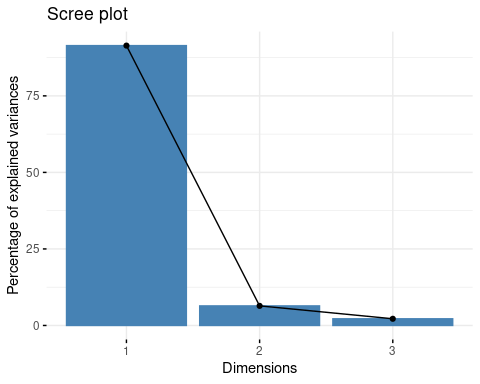
Now we are going to do Principle Component Analysis(PCA) with a variance threshold of 95% and plot out the scree plot.

There are 2 methods for PCA - spectral decomposition and singular value decomposition. We will try both methods and compare the results

# PCA spectral decomposition method  
pca.spectral <- princomp(select(data,TV,radio,newspaper))  
summary(pca.spectral)

## Importance of components:  
## Comp.1 Comp.2 Comp.3  
## Standard deviation 85.6529445 22.66045549 13.24516437  
## Proportion of Variance 0.9141558 0.06398422 0.02186001  
## Cumulative Proportion 0.9141558 0.97813999 1.00000000

# Scree plot  
fviz\_eig(pca.spectral)



# Obtaining the eigenvalues and var explained  
get\_eig(pca.spectral)

## eigenvalue variance.percent cumulative.variance.percent  
## Dim.1 7336.4269 91.415577 91.41558  
## Dim.2 513.4962 6.398422 97.81400  
## Dim.3 175.4344 2.186001 100.00000

# The contributions of each features in each component   
pca.spectral.var <- get\_pca\_var(pca.spectral)  
print(round(pca.spectral.var$contrib,5))

## Dim.1 Dim.2 Dim.3  
## TV 99.96590 0.03266 0.00145  
## radio 0.01003 12.77078 87.21919  
## newspaper 0.02408 87.19656 12.77936

# The coefficients of each features in each component  
pca.spectral$loadings

##   
## Loadings:  
## Comp.1 Comp.2 Comp.3  
## TV 1.000   
## radio -0.357 -0.934  
## newspaper -0.934 0.357  
##   
## Comp.1 Comp.2 Comp.3  
## SS loadings 1.000 1.000 1.000  
## Proportion Var 0.333 0.333 0.333  
## Cumulative Var 0.333 0.667 1.000

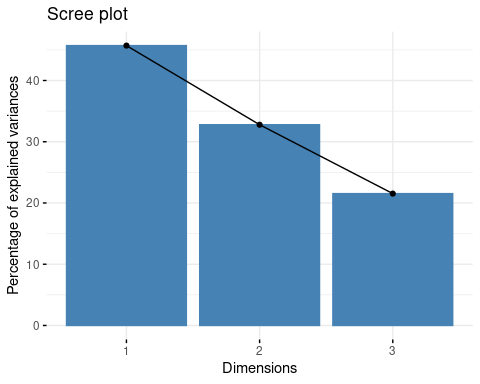
With the spectral decomposition approach, we can see that the 1st component represents roughly 91.4% of the variance, 2nd component represents roughly 6.4% of the the variance, and the 3rd component represents roughly 2.2% of the variance. Hence, with a variance threshold of 95%, 2 components will be selected, and roughly 97.8% of the variance are explained by the 2 components.

Next is the PCA with SVD method.

# PCA SVD method  
pca.svd <- prcomp(select(data,TV,radio,newspaper),   
 scale. = TRUE, center = TRUE)  
summary(pca.svd)

## Importance of components:  
## PC1 PC2 PC3  
## Standard deviation 1.171 0.9916 0.8037  
## Proportion of Variance 0.457 0.3277 0.2153  
## Cumulative Proportion 0.457 0.7847 1.0000

# Scree plot  
fviz\_eig(pca.svd)



# Obtaining the eigenvalues and var explained  
get\_eig(pca.svd)

## eigenvalue variance.percent cumulative.variance.percent  
## Dim.1 1.3708525 45.69508 45.69508  
## Dim.2 0.9832561 32.77520 78.47029  
## Dim.3 0.6458914 21.52971 100.00000

# The contribution of each features in each component   
pca.svd.var <- get\_pca\_var(pca.svd)  
print(round(pca.spectral.var$contrib,5))

## Dim.1 Dim.2 Dim.3  
## TV 99.96590 0.03266 0.00145  
## radio 0.01003 12.77078 87.21919  
## newspaper 0.02408 87.19656 12.77936

# The coefficients of each features in each component  
pca.svd$rotation

## PC1 PC2 PC3  
## TV 0.2078739 -0.9781484 0.003765898  
## radio 0.6913967 0.1496553 0.706805372  
## newspaper 0.6919241 0.1443227 -0.707398038

However, with the singular value decomposition approach, we can see that the 1st component represents roughly 45.7% of the variance, 2nd component represents roughly 32.8% of the the variance, and the 3rd component represents roughly 21.5% of the variance. Hence, with a variance threshold of 95%, all components will be selected, and 100% of the variance are explained by the 3 components.

**Note:** When I compare my result with my friends that compute using MATLAB, I realise that MATLAB is using Spectral Decomposition method, where 2 components will be selected with 95% variance threshold.