Thursday Exam 1

### **Background**

You are working as a marketing analyst for a professional basketball team seeking to improve **ticket sales** through more effective promotional strategies. Over the course of a season, the team has run 100 marketing campaigns and recorded a variety of data points for each. These campaigns span digital outreach, physical advertising, and fan engagement tactics. Your job is to analyze the data and develop a regression model that helps predict total sales (sales) based on these different forms of marketing activity.

The variables collected for each campaign include: the number of email\_clicks generated from newsletters, the number of social\_media\_mentions observed across platforms, the amount of ad\_spend used on digital platforms, the number of merch\_giveaways distributed during events, the estimated billboard\_exposure in terms of views, and the number of tv\_spots aired during peak hours. Using this dataset, you will build and compare several linear regression models to assess which combinations of marketing efforts most effectively predict ticket sales. This data was created based on simulated data.

### **Data Generation**

Warning: package 'broom' was built under R version 4.3.1

Warning: package 'lmtest' was built under R version 4.3.1

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

Loading required package: carData

Attaching package: 'car'

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

ticket\_sales email\_clicks social\_media\_mentions ad\_spend  
ticket\_sales 1.00 0.50 0.23 0.64  
email\_clicks 0.50 1.00 0.09 0.31  
social\_media\_mentions 0.23 0.09 1.00 0.58  
ad\_spend 0.64 0.31 0.58 1.00  
merch\_giveaways 0.37 0.20 0.12 0.23  
billboard\_exposure 0.23 0.09 0.02 0.11  
tv\_spots 0.37 0.17 0.12 0.22  
 merch\_giveaways billboard\_exposure tv\_spots  
ticket\_sales 0.37 0.23 0.37  
email\_clicks 0.20 0.09 0.17  
social\_media\_mentions 0.12 0.02 0.12  
ad\_spend 0.23 0.11 0.22  
merch\_giveaways 1.00 0.10 0.11  
billboard\_exposure 0.10 1.00 0.05  
tv\_spots 0.11 0.05 1.00

### **Model Building**

## Model 1  
model1 <- lm(ticket\_sales ~ email\_clicks + social\_media\_mentions, data = marketing\_df)  
  
## Model 2  
model2 <- lm(ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend, data = marketing\_df)  
  
## Model 3  
model3 <- lm(ticket\_sales ~ merch\_giveaways + billboard\_exposure + tv\_spots, data = marketing\_df)

model\_1\_summary <- summarize\_reg\_model(model1,"ticket\_sales ~ email\_clicks + social\_media\_mentions")  
model\_1\_summary <- summarize\_reg\_model(model1,"ticket\_sales ~ email\_clicks + social\_media\_mentions")  
model\_2\_summary <- summarize\_reg\_model(model2,"ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend")  
model\_3\_summary <- summarize\_reg\_model(model3,"ticket\_sales ~ merch\_giveaways + billboard\_exposure + tv\_spots")  
bind\_rows(  
 model\_1\_summary,  
 model\_2\_summary,  
 model\_3\_summary  
)

type RSS RSE  
1 ticket\_sales ~ email\_clicks + social\_media\_mentions 7274699 85.42  
2 ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend 4769075 69.20  
3 ticket\_sales ~ merch\_giveaways + billboard\_exposure + tv\_spots 7399156 86.19  
 R2 Adj\_R2 AIC BIC  
1 0.29 0.29 11738.03 11757.67  
2 0.53 0.53 11317.78 11342.32  
3 0.28 0.28 11757.00 11781.54

An employee at the company suggested that model 2 should be used for predicting ticket sales.

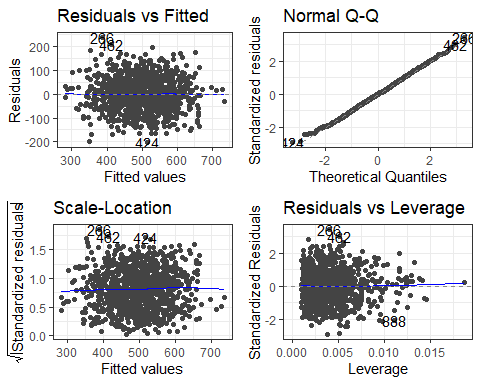
summary(model2)

Call:  
lm(formula = ticket\_sales ~ email\_clicks + social\_media\_mentions +   
 ad\_spend, data = marketing\_df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-201.427 -45.660 -0.031 47.048 236.109   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 216.34104 8.60595 25.14 < 2e-16 \*\*\*  
email\_clicks 0.27999 0.01986 14.10 < 2e-16 \*\*\*  
social\_media\_mentions -0.13702 0.02221 -6.17 9.92e-10 \*\*\*  
ad\_spend 0.51592 0.02255 22.88 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 69.2 on 996 degrees of freedom  
Multiple R-squared: 0.5344, Adjusted R-squared: 0.533   
F-statistic: 381.1 on 3 and 996 DF, p-value: < 2.2e-16

anova(model1,model2)

Analysis of Variance Table  
  
Model 1: ticket\_sales ~ email\_clicks + social\_media\_mentions  
Model 2: ticket\_sales ~ email\_clicks + social\_media\_mentions + ad\_spend  
 Res.Df RSS Df Sum of Sq F Pr(>F)   
1 997 7274699   
2 996 4769075 1 2505623 523.29 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Perform residual analysis plot  
autoplot(model2) + ## from ggfortify  
 theme\_bw()



resettest(model2) # From lmtest package

RESET test  
  
data: model2  
RESET = 0.16078, df1 = 2, df2 = 994, p-value = 0.8515

dwtest(model2) # From lmtest package

Durbin-Watson test  
  
data: model2  
DW = 1.9668, p-value = 0.2995  
alternative hypothesis: true autocorrelation is greater than 0

shapiro.test(resid(model2)) # Base R

Shapiro-Wilk normality test  
  
data: resid(model2)  
W = 0.99881, p-value = 0.7639

bptest(model2) # From lmtest package

studentized Breusch-Pagan test  
  
data: model2  
BP = 1.3747, df = 3, p-value = 0.7115

**Model 2: VIF**

vif(model2) ## From car package

email\_clicks social\_media\_mentions ad\_spend   
 1.119391 1.517900 1.663753

**Model 2: Partial Plots**

avPlots(model2) ## From car package

