Advertising Impact on Sales

Objective

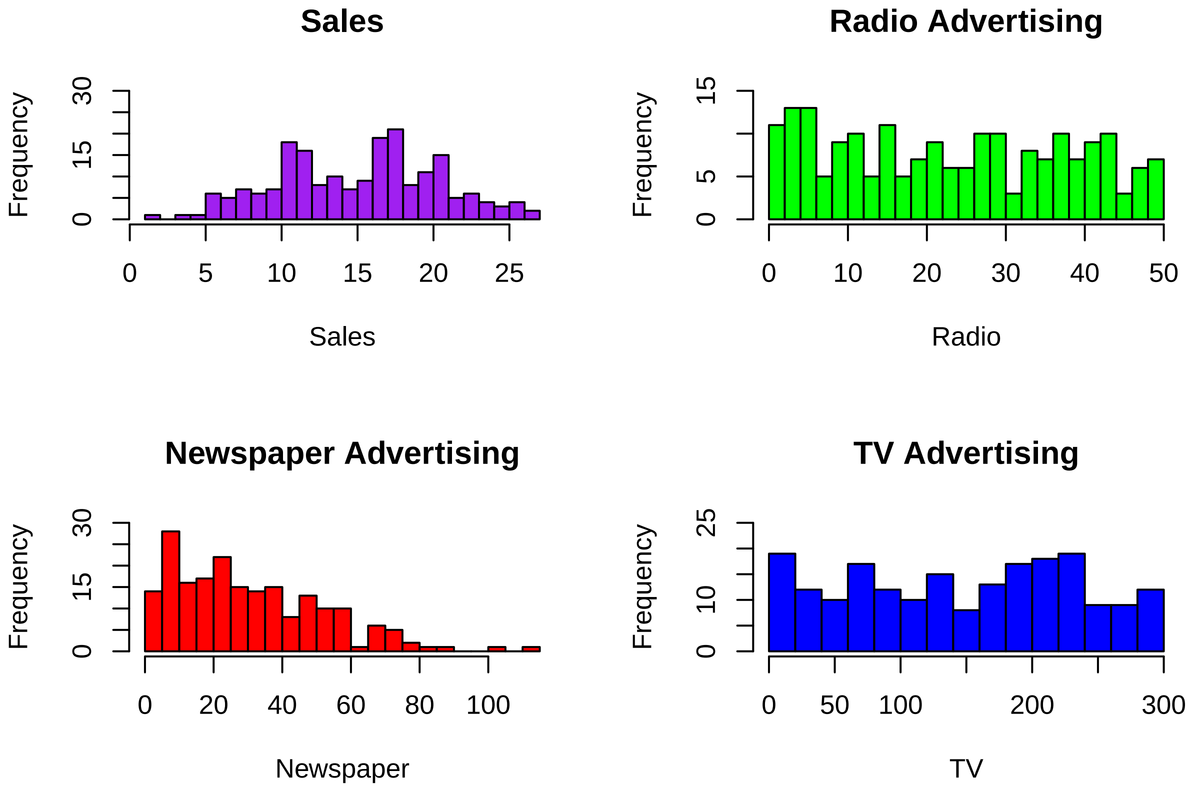
To determine how different types of advertising (TV, Radio, and Newspaper) influence sales, and to build the most effective linear regression model for predicting sales based on these advertising expenditures.

Data Description

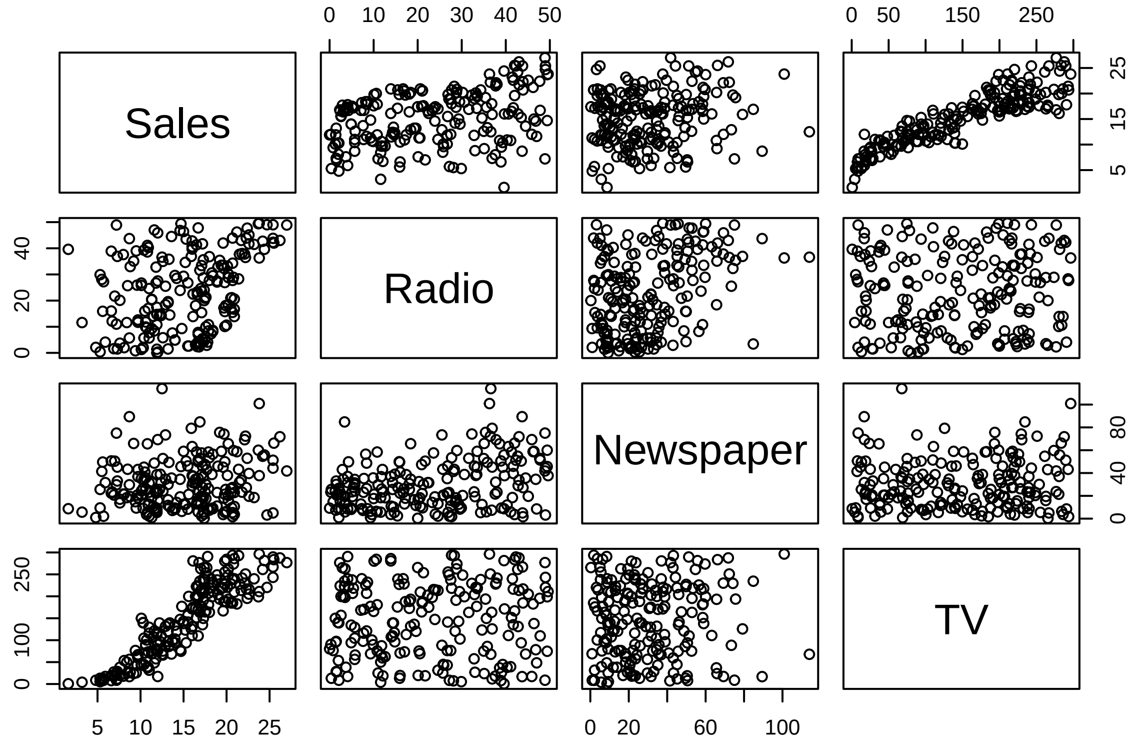
```r
advertising <- read.csv("advertising.csv")
attach(advertising)
# Calculate mean and standard deviation for each variable
mean_tv <- mean(TV)
std_tv <- sd(TV)

mean_radio <- mean(Radio)
std_radio <- sd(Radio)

mean_newspaper <- mean(Newspaper)
std_newspaper <- sd(Newspaper)

mean_sales <- mean(Sales)
std_sales <- sd(Sales)

summary_table_5x3 <- data.frame(
  Variable = c("Sales","Radio", "Newspaper","TV"),
  Mean = c(mean_sales, mean_radio, mean_newspaper, mean_tv),
  Std_Dev = c(std_sales,std_radio, std_newspaper,std_tv)
)
print(summary_table_5x3)
```

```
##    Variable     Mean   Std_Dev
## 1     Sales  15.1305  5.283892
## 2     Radio  23.2640 14.846809
## 3 Newspaper  30.5540 21.778621
## 4        TV 147.0425 85.854236
```

```r
#Histograms
par(mfrow=c(2,2))
hist(Sales,main="Sales", xlab="Sales", ylab="Frequency", col="purple", border="black",breaks=20,ylim=c(
hist(Radio,main="Radio Advertising", xlab="Radio", ylab="Frequency", col="green", border="black",breaks=
hist(Newspaper,main="Newspaper Advertising", xlab="Newspaper", ylab="Frequency", col="red", border="bla
hist(TV, main="TV Advertising", xlab="TV", ylab="Frequency", col="blue", border="black",breaks=20, ylim=
```



```
#Scatter plot of model
model <- lm(Sales~Radio + Newspaper + TV, data=advertising)
pairs(Sales~Radio + Newspaper + TV, data=advertising)
```



From the scatterplot matrix, it is evident that the TV advertising budget shows a strong linear relationship with Sales, whereas Radio and Newspaper do not display any clear linear trends with the response variable. The variances of all three predictors appear to be random and relatively constant, indicating no major issues with heteroscedasticity. Additionally, there are a few noticeable outliers in the Sales vs. Newspaper plot. Importantly, there is no visible multicollinearity among the predictors, as the variables are not strongly correlated with each other.

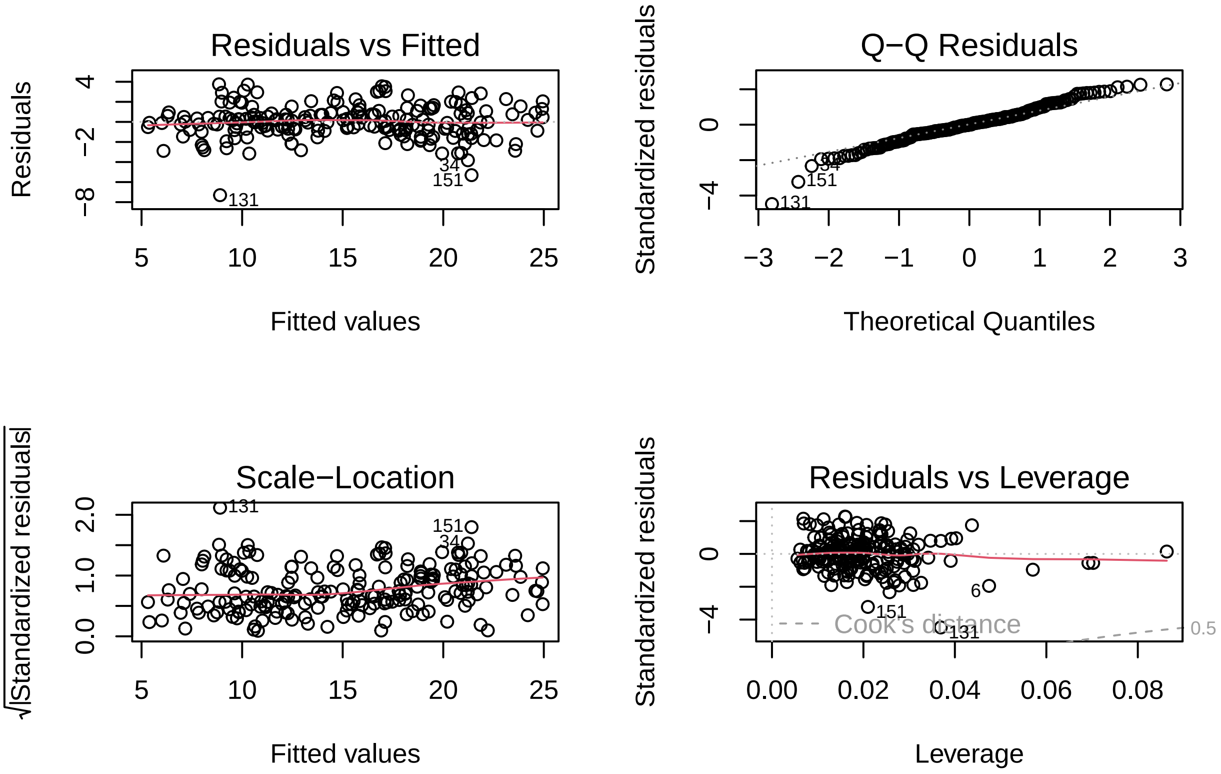
Results and Interpretation

```
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ Radio + Newspaper + TV, data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3034 -0.8244 -0.0008  0.8976  3.7473
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.6251241  0.3075012  15.041   <2e-16 ***
## Radio       0.1070012  0.0084896  12.604   <2e-16 ***
## Newspaper   0.0003357  0.0057881   0.058    0.954
## TV          0.0544458  0.0013752  39.592   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.662 on 196 degrees of freedom
## Multiple R-squared:  0.9026, Adjusted R-squared:  0.9011
## F-statistic: 605.4 on 3 and 196 DF,  p-value: < 2.2e-16
```

The multiple linear regression model Sales ~ TV + Radio + Newspaper was fit to the data to assess the impact of different advertising channels on sales. Based on the summary output, TV and Radio were found to be highly significant predictors (p-values < 2e-16), while Newspaper had no significant contribution (p = 0.954). The model has a high R-squared of 0.9026, indicating that approximately 90% of the variance in Sales is explained by the three predictors. The F-statistic is also highly significant, supporting the overall model fit.

```r
par(mfrow=c(2,2))
#Diagnostic Plot
plot(model)
```



Diagnostic plots indicate that the model meets the assumption of linearity, with no strong curvature in the residuals vs. fitted plot. The Q-Q plot suggests that the residuals are approximately normally distributed, although a few outliers are present. The scale-location plot confirms that the variance of the error term is roughly constant, though there's a slight trend that could be improved. Lastly, the residuals vs. leverage plot shows that some data points may be influential, but most observations fall within an acceptable leverage range.

```
#Transfor/Filter Data
library(car)
```

```
## Loading required package: carData
```

```
filteredmodel <- advertising + 0.000001
detach(advertising)
attach(filteredmodel)
#because Radio has 1 zero value data
```

```
#Box Cox Transformation
summary(powerTransform(cbind(Sales,Radio,Newspaper,TV)~1,filteredmodel))
```

```
## bcPower Transformations to Multinormality
##           Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Sales        0.6598        0.50       0.4909       0.8287
## Radio        0.7240        0.72       0.5833       0.8648
## Newspaper    0.4611        0.50       0.3223       0.5999
## TV           0.4470        0.50       0.3361       0.5578
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                 LRT df       pval
## LR test, lambda = (0 0 0 0) 473.3372  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                 LRT df       pval
## LR test, lambda = (1 1 1 1) 134.3371  4 < 2.22e-16
```

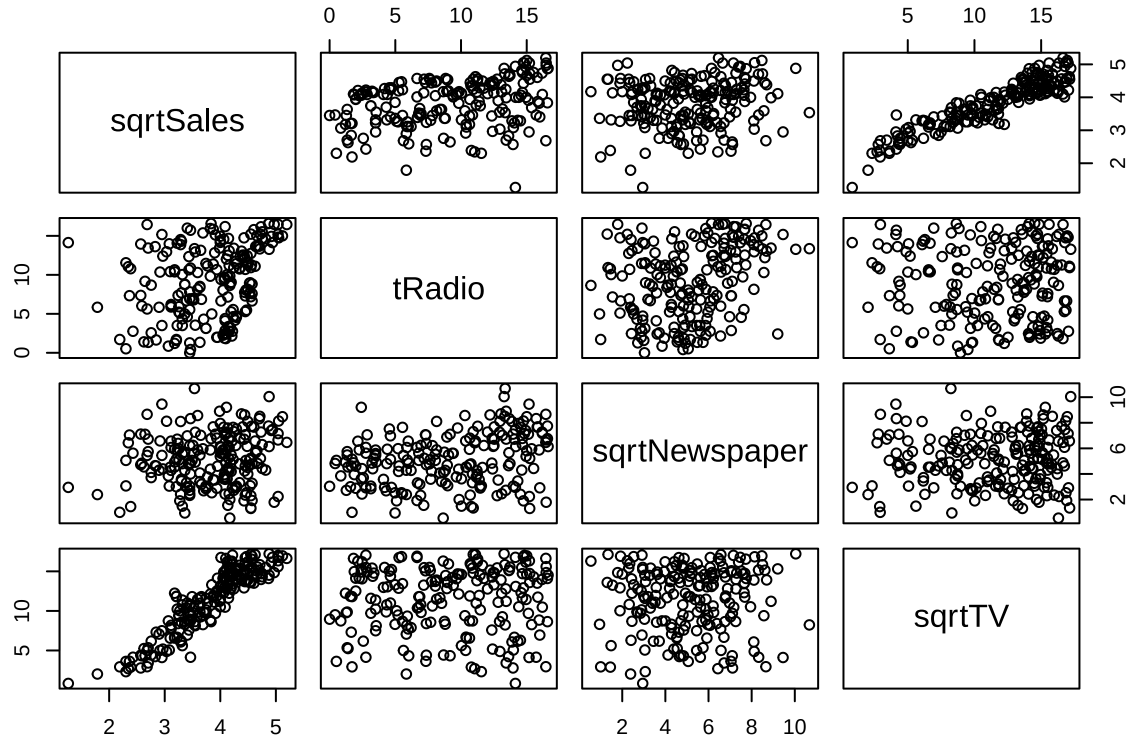
```
#transform y and x
sqrtSales <- sqrt(Sales)
sqrtNewspaper <- sqrt(Newspaper)
sqrtTV <- sqrt(TV)
tRadio <- Radio^(0.72)
```

To improve model assumptions, a Box-Cox transformation was applied to the variables to assess whether transformation would improve multivariate normality. The estimated power transformations suggested that all variables should be approximately square root or log transformed, with optimal lamda values around 0.5 for TV and Newspaper, 0.72 for Radio, and 0.66 for Sales. The likelihood ratio tests strongly rejected both the null hypotheses that all variables require no transformation, and all should be log-transformed, with p-values < 2.2e-16 in both cases.

```
transformedmodel <- lm(sqrtSales~tRadio+sqrtNewspaper+sqrtTV)
summary(transformedmodel)
```

```
##
## Call:
## lm(formula = sqrtSales ~ tRadio + sqrtNewspaper + sqrtTV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08208 -0.11616 -0.00294  0.10860  0.49657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.593469   0.055526  28.698   <2e-16 ***
## tRadio       0.042869   0.003128  13.705   <2e-16 ***
## sqrtNewspaper 0.004941   0.007199   0.686    0.493
## sqrtTV       0.158886   0.003382  46.986   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1971 on 196 degrees of freedom
## Multiple R-squared:  0.9266, Adjusted R-squared:  0.9255
## F-statistic:   825 on 3 and 196 DF,  p-value: < 2.2e-16
```

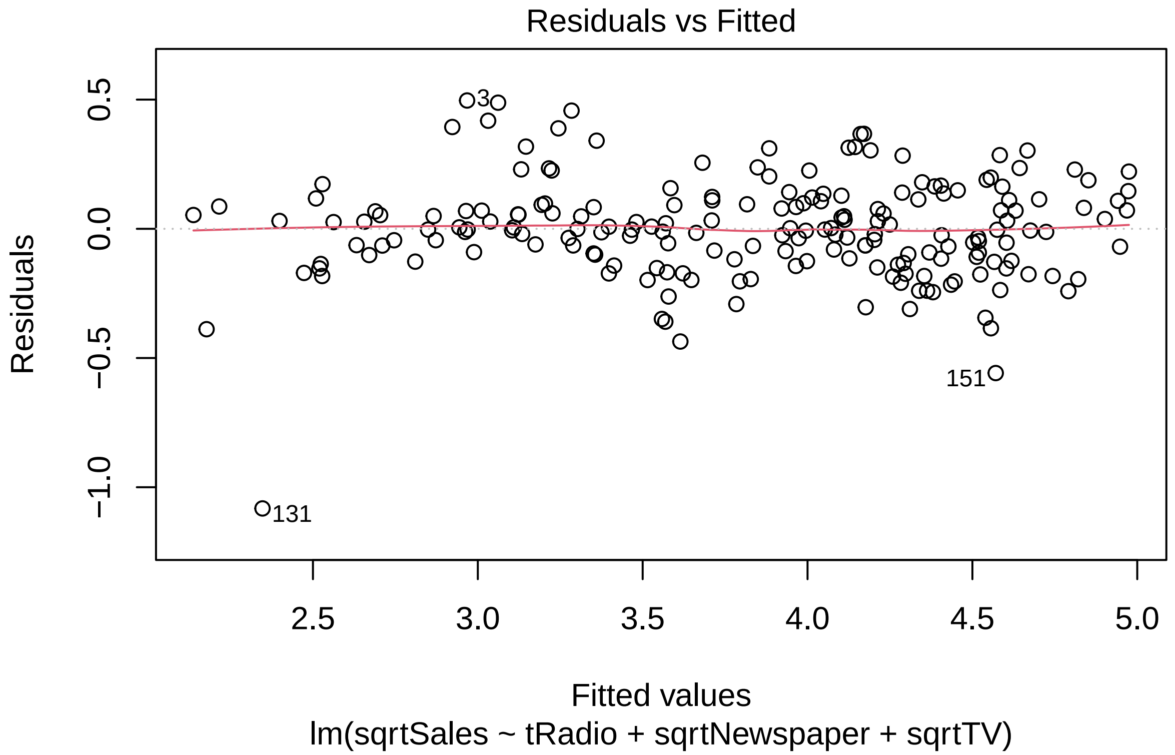
```
pairs(sqrtSales~tRadio+sqrtNewspaper+sqrtTV)
```

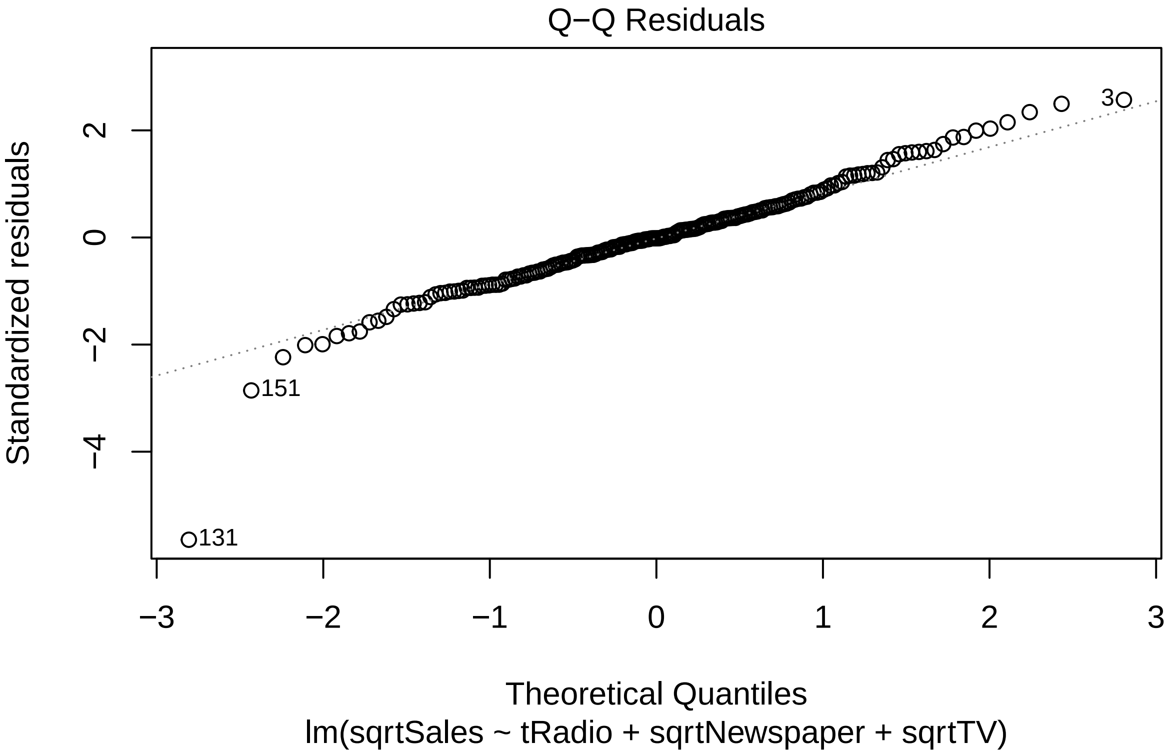


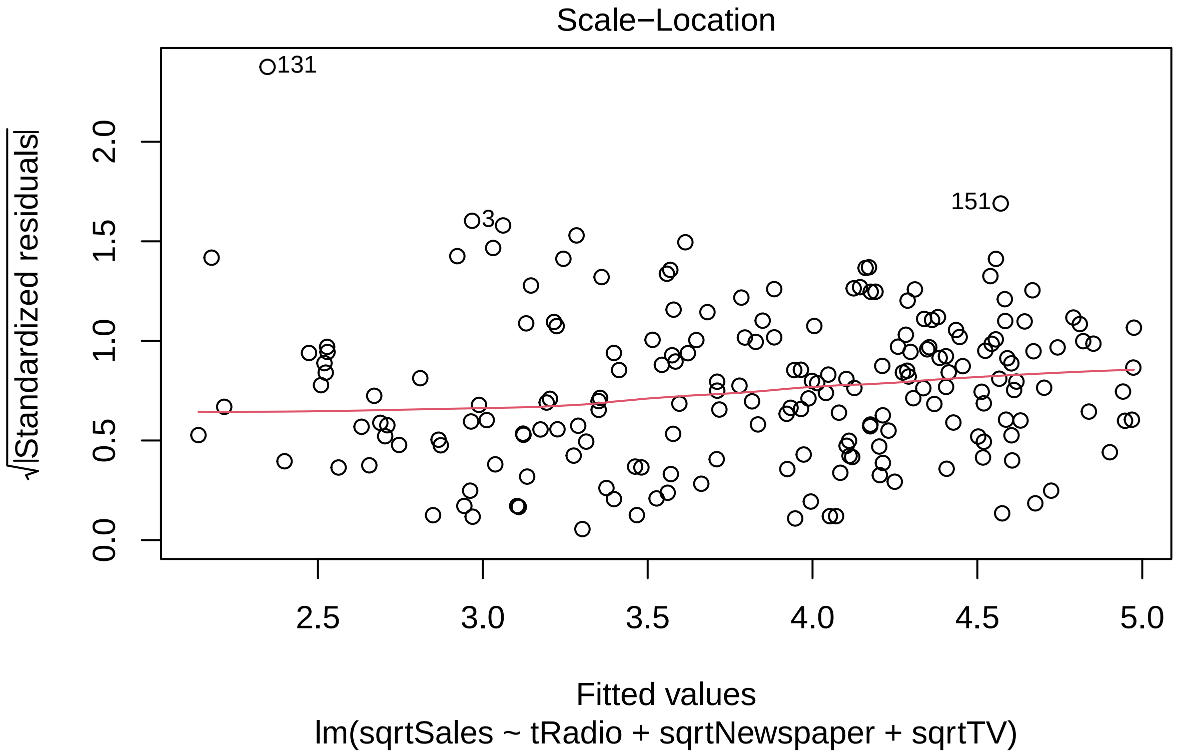
```
par(mfrow=c(2,2))
```

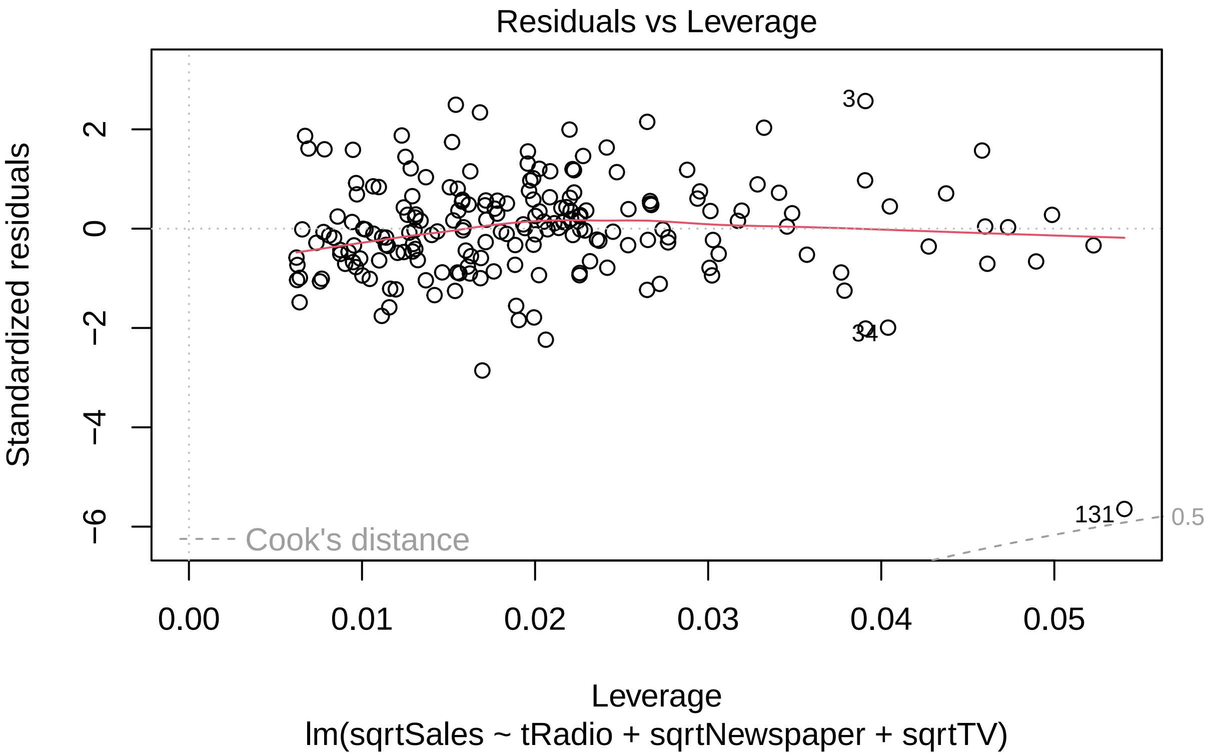
This transformed model yielded a higher adjusted R-squared of 0.9255, indicating improved explanatory power compared to the original model. Both tRadio and sqrt(TV) remained statistically significant predictors (p < 2e-16), while sqrt(Newspaper) remained insignificant, aligning with earlier findings.

```
#Diagnostic Plot for transformed model
plot(transformedmodel)
```



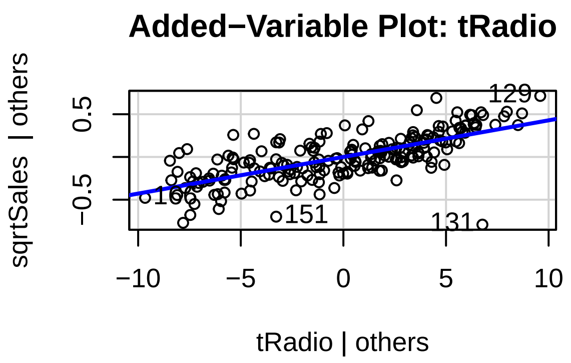

Q−Q Residuals
Standardized residuals
2
0
−2
−4
−3
−2
−1
0
1
2
3
3
151
131
Theoretical Quantiles
lm(sqrtSales ~ tRadio + sqrtNewspaper + sqrtTV)

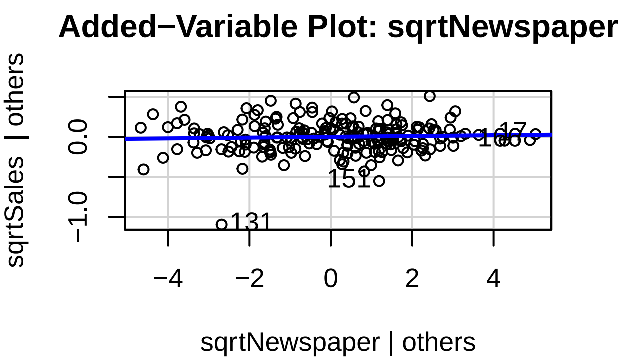

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
131
3
151
2.5
3.0
3.5
4.0
4.5
5.0
Fitted values
lm(sqrtSales ~ tRadio + sqrtNewspaper + sqrtTV)

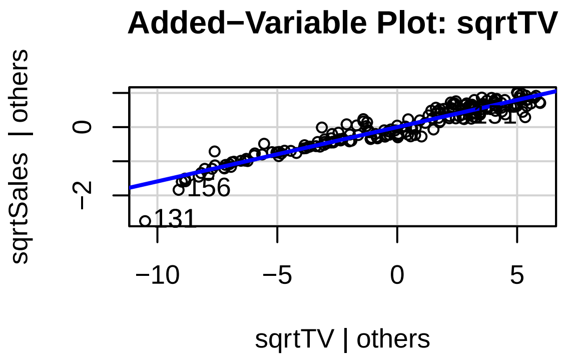


Diagnostic plots confirm improved model fit: residuals appear more linear and normally distributed, with constant variance and no major outliers. Overall, the transformed model better satisfies regression assumptions and improves predictive accuracy.

```r
#Check added variable plot for each predictor
par(mfrow=c(2,2))
avPlot(transformedmodel,variable = tRadio, ask=FALSE)
avPlot(transformedmodel,variable = sqrtNewspaper, ask=FALSE)
avPlot(transformedmodel,variable = sqrtTV, ask=FALSE)
```







The added-variable plots show that tRadio and sqrtTV have strong partial relationships with the response variable sqrtSales, confirming their importance in the model. The plot for sqrtNewspaper, however, shows no noticeable trend, suggesting it does not contribute significantly once other variables are accounted for.

```
#check for collinearity
vif(transformedmodel)
```

```
##        tRadio sqrtNewspaper        sqrtTV
##      1.102518      1.103186      1.002233
```

The Variance Inflation Factors (VIFs) for all predictors are close to 1, indicating no multicollinearity. This validates that the predictors are independent and stable in the model.

```
#Use Forward Stepwise regression for variable selection
#since newspapers var does not have sign slope
mint <- lm(sqrtSales~1,data=advertising)
forwardAIC <- step(mint,scope=list(lower=~1,
upper=~tRadio+sqrtNewspaper+sqrtTV),
direction="forward",data=advertising)
```

```
## Start:  AIC=-129.2
## sqrtSales ~ 1
##
##                 Df Sum of Sq    RSS     AIC
## + sqrtTV         1    87.866 15.917 -502.18
## + tRadio         1    10.192  93.591 -147.88
## + sqrtNewspaper  1     1.938 101.845 -130.97
## <none>                       103.783 -129.20
##
## Step:  AIC=-502.18
## sqrtSales ~ sqrtTV
##
##                 Df Sum of Sq    RSS     AIC
## + tRadio         1    8.2834  7.634 -647.14
## + sqrtNewspaper  1    1.0033 14.914 -513.20
## <none>                       15.917 -502.18
##
## Step:  AIC=-647.14
## sqrtSales ~ sqrtTV + tRadio
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       7.6340 -647.14
## + sqrtNewspaper  1  0.018305 7.6157 -645.62
```

Using AIC-based forward selection, the procedure initially identified sqrtTV as the strongest single predictor, followed by tRadio. Although sqrtNewspaper was tested, it did not improve the model and was ultimately excluded. The final model selected was sqrtSales ~ sqrtTV + tRadio.

```r
reducedmodel <- lm(sqrtSales~sqrtTV+tRadio)
summary(reducedmodel)
```

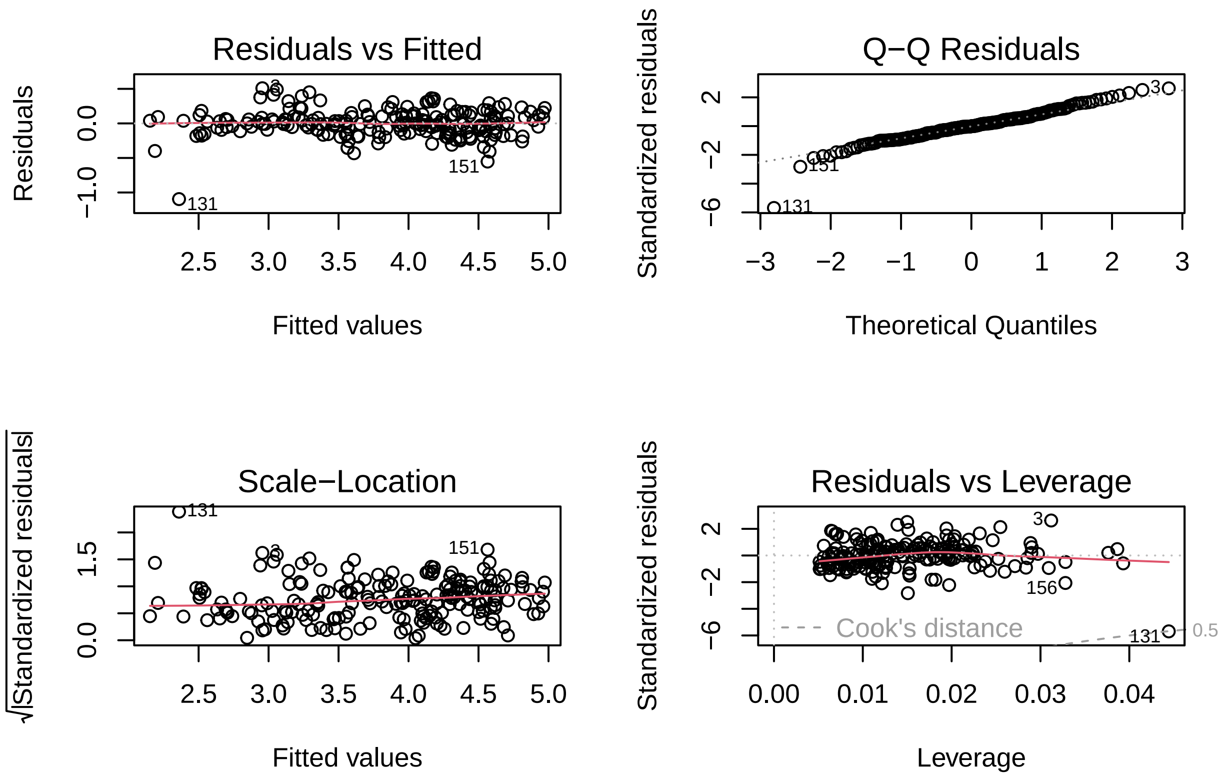
```
##
## Call:
## lm(formula = sqrtSales ~ sqrtTV + tRadio)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.09534 -0.11320 -0.00162  0.10627  0.50858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.612038   0.048425   33.29   <2e-16 ***
## sqrtTV      0.158963   0.003375   47.10   <2e-16 ***
## tRadio      0.043520   0.002977   14.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1969 on 197 degrees of freedom
## Multiple R-squared:  0.9264, Adjusted R-squared:  0.9257
## F-statistic:  1241 on 2 and 197 DF,  p-value: < 2.2e-16
```

```r
#Partial F-test to compare the transformed and reduced model
anova(reducedmodel,transformedmodel)
```

```
## Analysis of Variance Table
##
## Model 1: sqrtSales ~ sqrtTV + tRadio
## Model 2: sqrtSales ~ tRadio + sqrtNewspaper + sqrtTV
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    197 7.6340
## 2    196 7.6157  1  0.018305 0.4711 0.4933
```

A partial F-test was then used to compare the reduced model against the full model. The p-value of 0.4933 was well above 0.05, leading to the conclusion that the simpler model was sufficient.

```
#Diagnostic Plot for final model
par(mfrow=c(2,2))
plot(reducedmodel)
```



Residual diagnostics for the final model confirmed that assumptions of linearity, constant variance, and normality held reasonably well, with only minor outliers.