# STATS 101A - Final Project Report: Predicting Sales from Ad Budgets

## Introduction

The purpose of this research project is to develop a predictive model of product sales to study how advertising budgets for different mediums impacted sales revenue. The dataset utilized in this study was obtained from Kaggle, containing two hundred observations and four variables to analyze: Sales, TV Ad Budget, Radio Ad Budget, and Newspaper Ad Budget:

- sales: Sales Revenue in millions of dollars
- tvAd: Budget: Budget for TV Advertisements in thousands of dollars
- radioAd: Budget for Radio Advertisements in thousands of dollars
- newspaperAd: Budget for Newspaper Advertisements in thousands of dollars

Being able to discern which medium of advertising results in the highest increase in sales revenue is an important topic for any sales team, so we let sales be the response variable and tvAd, radioAd, and newspaperAd be the predictor variables. All the analysis for this study was conducted using R. Multiple linear regression was used to model the relationship between the predictor and response variables since the initial scatterplot matrix indicated some linear relationship between the variables.

This paper first looks over general summary statistics of all the variables. The generalized full model is first used to fit all the untransformed variables. Transformation and variable selection were then analyzed to find a model that may better fit the trend in variables. Throughout this process, each developed model was analyzed for model assumption satisfaction and predictor significance. After conducting this series of investigations, a final model is found.

## Data Description

We start by first examining the summary statistics. We can see that the response variable (sales) is normally distributed, the TV ad and radio ad budgets are somewhat uniformly distributed, and the newspaper ad budget is right skewed.

| variable | mean | sd |
|---|---|---|
| sales | 14.022 | 5.217 |
| tvAd | 147.042 | 85.854 |
| radioAd | 23.264 | 14.847 |
| newspaperAd | 30.554 | 21.779 |

*Table 1. Variable means and standard deviations*

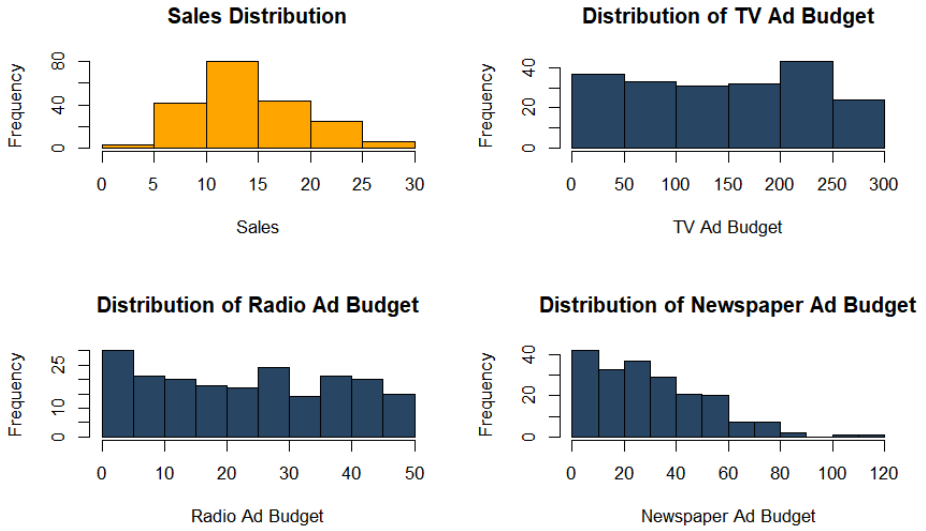


Figure 2. Distribution of variables

Table 1 summarizes each variable's mean and standard deviation. We can see that the highest budget is spent on TV ads, while the lowest is spent on radio ads. TV ad budget also has the greatest standard deviation.

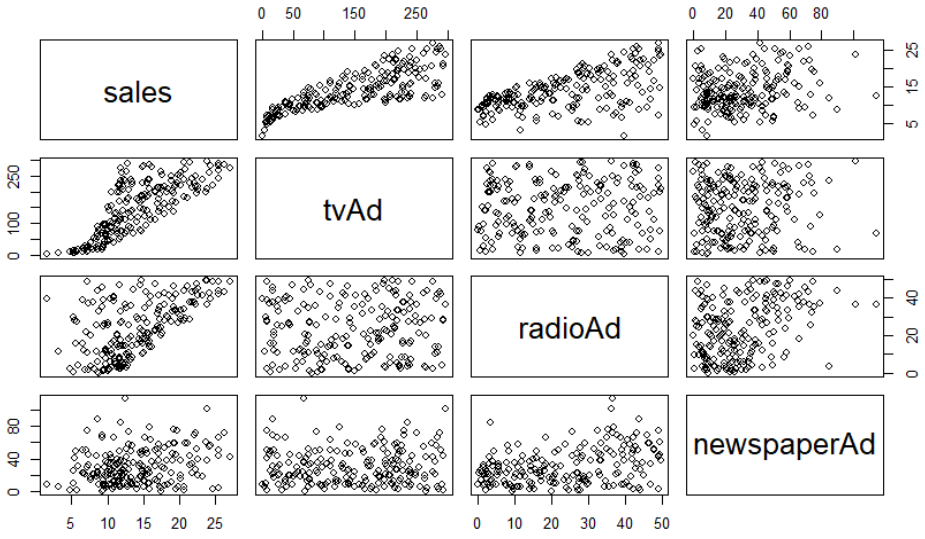


*Figure 1. Scatterplot matrix of variables*

Looking at the scatter plot matrix in Figure 1, it appears sales has a positive linear relationship with both tvAd and radioAd. However, there appears to be heteroscedasticity in the relationship between sales and tvAd. There doesn't appear to be an apparent relationship between sales and newspaperAd, which will be further investigated through later analysis. Finally, there does not seem to be a relationship between the three predictor variables, which seem to be somewhat random. This seems to suggest a lack of multicollinearity, which is good.

Given the apparent linear relationship between some of the predictor variables and the response variable, we chose to begin our analysis by fitting the data with a multiple linear regression model.

### Results and Interpretation

Our first candidate model is the full model with untransformed data.

```
Call:
lm(formula = sales ~ tvAd + radioAd + newspaperAd)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8277 -0.8908  0.2418  1.1893  2.8292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.938889   0.311908   9.422   <2e-16 ***
tvAd         0.045765   0.001395  32.809   <2e-16 ***
radioAd      0.188530   0.008611  21.893   <2e-16 ***
newspaperAd -0.001037   0.005871  -0.177     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.686 on 196 degrees of freedom
Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

*Figure 2. Summary Statistics for Candidate Model 1*

From Figure 2 we derive the linear regression equation:

$$\hat{sales} = 2.939 + 0.046\,tvAd + 0.189\,radioAd - 0.001\,newspaperAd$$

We can see that although tvAd and radioAd have significant p-values for their coefficients, newspaperAd does not, indicating that variable selection may be required later. In addition to the coefficients, the summary also reports an $R^2$ value of 0.8972, indicating that 89.72% of the variance in sales is explained by the model. Finally, the summary shows that the overall F-test was significant with a p-value less than 0.05, rejecting the null hypothesis and giving significant evidence that at least one of the predictor coefficients is significant.

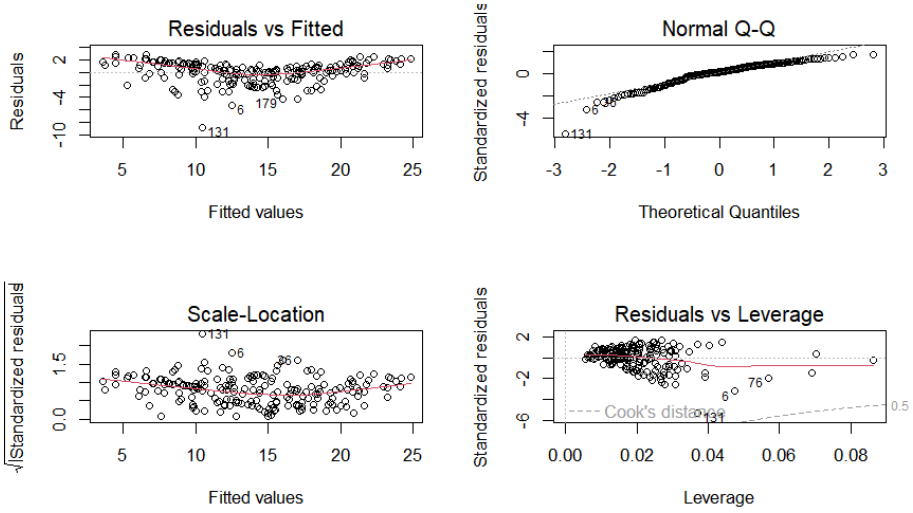


*Figure 3. Diagnostic Plots for Candidate Model 1*

Figure 3 shows that the full model mostly satisfies the model assumptions for multiple linear regression to some extent. The Residuals vs Fitted and Scale-Location plots do not appear to be a straight line, indicating some violation to the linearity and constant variance assumptions. There is also a slight left skew and heavy tail in the Normal Q-Q plot, indicating some violation of the normality assumption. Additionally, there are also some outliers shown in the Residuals vs. Leverage graph, which we found to be the points 6, 37, 76, 129, and 166.

To see if we could improve the diagnostic plots, we investigated if transformation is necessary through the Box-Cox method. The likelihood ratio test for all log transformations resulted in a p-value of $2.22*10^{-16}$, which is less than a significance value of $\alpha = 0.05$, indicating an all-log transformation is not recommended. Similarly, the likelihood ratio test that no transformation is needed resulted in a p-value of $2.22*10^{-16} < \alpha$, indicating that transformation is necessary. We then took the rounded power of each variable as suggested and fit the transformed variables to a new multiple linear regression model:

$$\hat{log}(sales) = 0.518 + 0.356\, log(tvAd) + 0.036\, (radioAd)^{0.76} - 0.002(\, newspaperAd)^{0.5}$$

**Note that observation 128 had an original radioAd budget of 0. Since log(0) is undefined, we changed this value to the arbitrarily small value of 0.00000000001 to ensure the transformation was defined.**

For this second candidate model, all predictors were again significant except for newspaperAd. Our second candidate model has a $R^2$ value of 0.9734, which is now higher than the first candidate model. The F-statistic of the model resulted in a significant p-value as well, suggesting that there is a significant linear association between the predictor and response variables of this model.

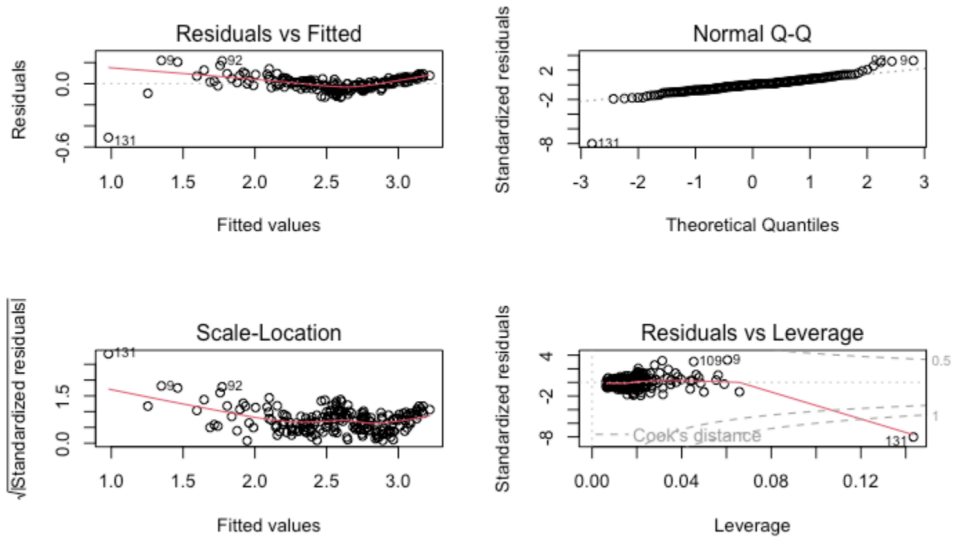


*Figure 4. Diagnostic Plots of Candidate Model 2 (Transformed)*

The diagnostic plots in Figure 4 show that many of the assumption violations still remain. The Residuals vs. Fitted and Scale-Location models show that the trend line is not horizontal, indicating violations to the linearity and constant variance assumptions. There is also a violation of the normality assumption, as the Normal Q-Q plot further does not follow a linear pattern. Finally, there also seems to be some leverage and influential points, although none of the points are considered bad leverage points.

However ,the diagnostic plots show that the influential point (based on Cook's Distance) 131 appears to be significant, showing up in all four graphs and heavily influencing the Residuals vs. Leverage trendline. Upon investigation, this point contained a particularly small sales and TV Ad budget, which deviated from the rest of the dataset. However, we do not have further information regarding this point and it is not considered a bad leverage point, so there is not enough justification to remove this point.

Based on the summary of the multiple linear regression for Model 2, we investigated variable selection due to the high p-value of the newspaperAd variable. We examined the added variable plots, in addition to conducting the $R^2$ adjusted, AIC, AIC corrected, BIC, and partial F-tests.

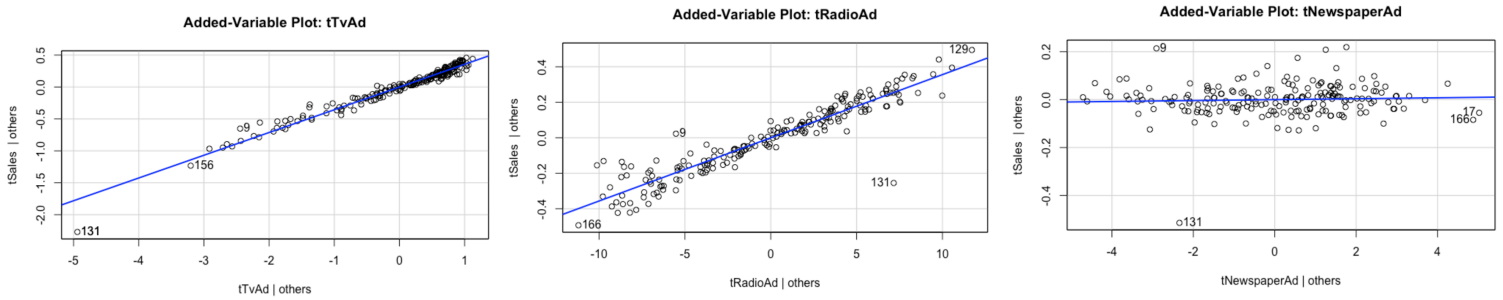


*Figure 6. Added Variable Plots*

| Size | $R^2$ Adjusted | AIC | AIC Corrected | BIC |
|---|---|---|---|---|
| 1 | 0.7408 | -620.4844 | -620.3631 | -613.8877 |
| **2** | **0.9726** | **-1068.8776** | **-1068.6746** | **-1058.9827** |
| 3 | 0.9725 | -1067.4894 | -1067.1832 | -1054.2961 |

*Table 3. $R^2$ Adjusted, AIC, AIC Corrected, BIC*

The added variable plots from Figure 6 shows that both tvAd and radioAd have significant slopes while newspaperAd does not. This is further supported by the $R^2_{adj}$, AIC, AIC corrected, and BIC tests, which all indicate that the model of size 2, which is the linear regression model of sales with the variables radioAd and tvAd, is the better model. Moreover, the partial F-test between the reduced and full model has a p-value of 0.4394 > $\alpha$ = 0.05, leading us to the conclusion that we fail to reject the null hypothesis and that the reduced model is a better fit for the data.

Thus, we arrive at our final model, which is transformed and reduced:

$$\hat{log}(sales) = 0.525 + 0.356\, log(tvAd) + 0.036(radioAd)^{0.76}$$

For our final model, all predictors are significant with p-values less than 0.05. Further, the adjusted $R^2$ is very high at 0.9726 and the F-statistic is significant with a p-value less than 0.05.

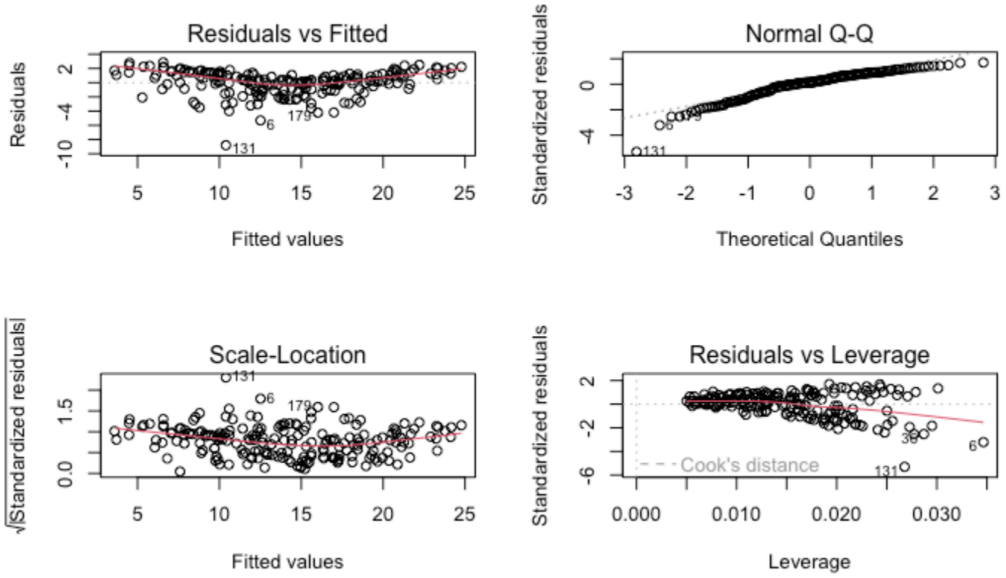


*Figure 7. Diagnostic Plots of Candidate Model 3 (Final)*

Checking our diagnostic plot for this final model, we can see that other than point 131, there aren't any other influential points. There still seems to be some non-constant variance as shown in the Residuals v.s. Fitted plot, which isn't great, indicating transformation did not fix it entirely. The Normal Q-Q plot is mostly alright, with a small left skewing but nothing terrible. Given the higher $R^2$ value of this model and the Likelihood Ratio Test suggesting to transform variables, we chose to stick with the transformed model, and leave out the variable newspaperAd since the added variable plots, $R^2_{adj}$, AIC, AIC corrected, BIC, and partial F-test indicate that it is not a significant predictor.

**Discussion**

Our final model indicates a 1% increase in the tvAd budget results in a 0.357% increase in the sales revenue. The effect of radioAd, however, is harder to interpret since its power transformation is not 0 or 1. All we can say is that radioAd has a small positive effect on sales revenue. Thus, our results indicate that increasing television ad budget is the most effective form of increasing sales. Increasing radio ad budget is also effective, to a lesser extent as indicated by the smaller slope, while increasing newspaper ad budget does not seem to be an effective form of increasing sales.

These findings are reasonable as the new digital age transitions consumers towards screens rather than traditional radio and newspapers. This is further applicable to real-world situations as companies can utilize these findings and proportion their ad budgets accordingly across different mediums in the most effective/efficient way in order to maximize sales.

The primary limitations of this model lie in the slight violations of model assumptions, specifically non-constant variance. Other transformations or additional features such as quadratic terms in the linear model could be experimented with to fix this limitation. Additionally, future research could investigate if adding other predictor variables, such as the industry of the product being sold or additional forms of advertising (i.e. social media), may help increase the level of accuracy and applicability of this model.