Simple Regression Analysis

Joseph Simonian

Oct 07, 2016

Abstract

We analyze the relationship between advertisement budget and unit sales of a particular product across a number of markets, and across a variety of advertisement types. To do this, we compute several multivariate regressions, each predicting product Sales based on advertisement budgets across a range of media. Based on our regression results, we conclude that a higher budget towards television and radio advertisements is correlated with higher sales of a particular product.

Introduction

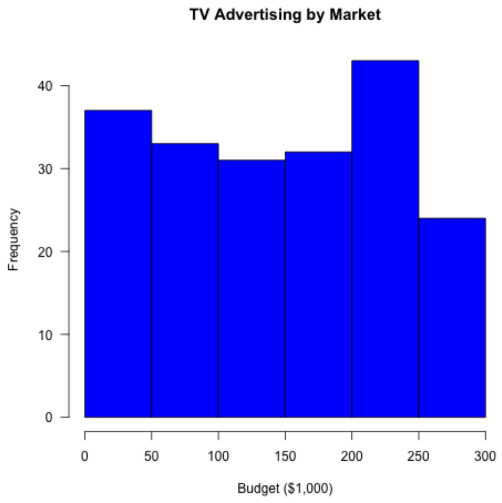
Paid advertising generates the profits behind many forms of media - most prominently television, but also websites, newspapers, radio, and other forms of media. Advertisers who wish to achieve the most "bang for their buck" will want to know which forms of media advertising tend to generate the most sales.

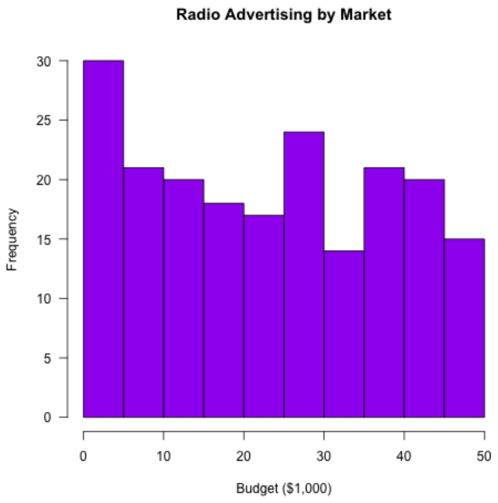
We compare advertisement budgets across a variety of media types in order to create a predictive model of sales based upon advertising budgets. In doing so, we additionally analyze which forms of advertising tend to increase sales the most. We thus produce two results: both a comparison of the effectiveness of different forms of advertising, and a predictive model for sales based on the amount budgeted towards certain forms of advertising.

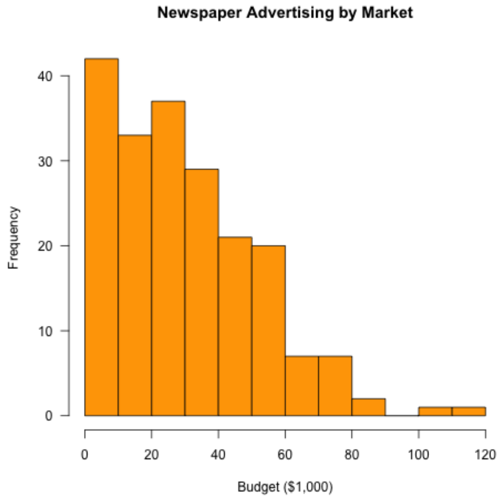
Data

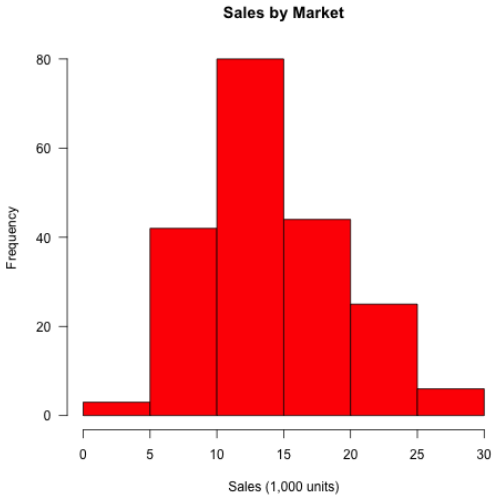
We use the `Advertising` dataset, which originally appeared in chapter 2.1 of "An Introduction to Statistical Learning", by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. This datset consists of the `Sales` (in thousands of units) of a product across 200 different markets, along with advertising budgets (in thousands of dollars) for the product in each market, across three types of media: `TV`, `Radio`, and `Newspaper`.

Histograms of the each advertising budget by region, and Sales by region, are shown below:






Newspaper Advertising by Market


Sales by Market

Methodology

We analyze the relationships between advertising budgets and Sales, across different markets. To do this, we use several multivariate linear models:

$$Sales = \beta_0 + \beta_1 * A_1 + \beta_2 * A_2 + ... + \mu$$

Where "$Sales$" represents the total sales (in thousands of units) for a region.

The coefficients A_1, A_2, etc. represent the advertising budgets in different forms of media.

β_0 represents the coefficient the amount of product that would have been sold regardless of advertising, and every other β_i represents the sensitivity of sales to advertising, that is, the increase in sales that can be expected from some increase in the advertising budget A_i.

μ represents error due to factors beyond advertising, and random noise.

After defining each model, we fit the parameters β_i via a multivariate least-squares regression.

Results

As we saw in homework 2, TV advertising budget was useful in predicting sales. First, we analyze univariate regressions on all three advertisement varieties below:

```
print(summary(TV_model)$coefficients)
```

```
##               Estimate  Std. Error  t value    Pr(>|t|)
## (Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
## TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

```
print(summary(Radio_model)$coefficients)
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
## Radio       0.2024958 0.02041131  9.920765 4.354966e-19
```

```
print(summary(Newspaper_model)$coefficients)
```

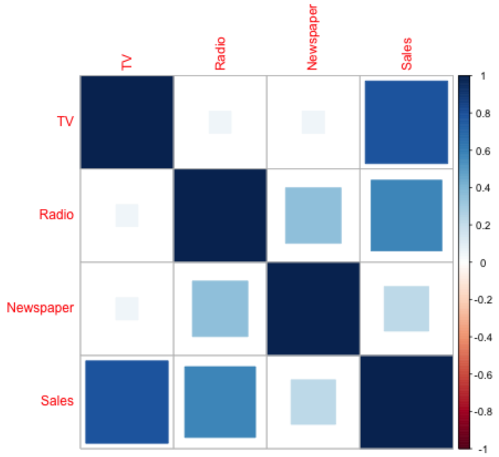
```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
## Newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

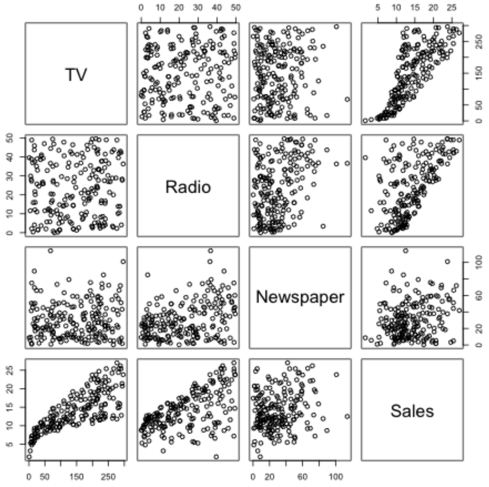
Based on these results, we can see that the increase in sales due to advertising is statistically significant for all three models, and that all three varieties of advertisements are positively correlated with sales. However, it could be the case that advertisers tend to spend money roughly equally on all three varieties of advertisements, and thus, this result is driven by a single form of advertising while the other forms are only correlated with sales due to their correlation with the driving factor.

Before considering any additional regression models, we can visualize the correlations between sales and different types of advertising. The values are shown below, both numerically and as a graphic. We also view scatterplots of how factors affect each other:

```
print(corr_matrix)
```

```
##                   TV      Radio  Newspaper     Sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## Radio     0.05480866 1.00000000 0.35410375 0.5762226
## Newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## Sales     0.78222442 0.57622257 0.22829903 1.0000000
```





It seems that Newspaper advertising budget is the least correlated with Sales. It is also fairly well correlated with Radio advertising budget - perhaps its effect on sales is really just due to its correlation with the predictive Radio advertising budget factor. To see if this is true, we use the backward selection method, first analyzing a multivariate regression on all three factors and then removing factors. The results of a regression including all three types of media budgets are shown below:

```
print(summary(Triple_model)$coefficients)
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## Radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## Newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

We can see that, in a multivariate regression, Newspaper advertising has a p-value above 0.05. Thus, we can conclude that it is not predictive when radio and television advertisements are accounted for.

We can check a model's goodness of fit by analyzing its R^2 and RSE value.s We compare the R^2 and RSE values of this triple model and a model that only factors in TV and Radio ads. The results can be seen below.

Values for 3-variable regression:

```
Quantity = c("Residual Standard Error", "R-squared", "F-statistic")
Value = c(residual_std_error(Triple_model),
          r_squared(Triple_model),
          f_statistic(Triple_model))
print(data.frame(Quantity, Value))
```

```
##                  Quantity       Value
## 1 Residual Standard Error   1.6855104
## 2               R-squared   0.8972106
## 3             F-statistic 570.2707037
```

Values for 2-variable regression of TV and Radio:

```
Quantity = c("Residual Standard Error", "R-squared", "F-statistic")
Value = c(residual_std_error(TV_Radio_model),
          r_squared(TV_Radio_model),
          f_statistic(TV_Radio_model))
print(data.frame(Quantity, Value))
```

```
##                  Quantity       Value
## 1 Residual Standard Error   1.6813609
## 2               R-squared   0.8971943
## 3             F-statistic 859.6177183
```

Since R^2 and RSE will always increase with the addition of more parameters, the fact that they are nearly the same after removing the Newspaper advertising budget parameter shows that the Newspaper parameter can be dropped from the model. So, our final model is one based only on TV and Radio advertising budgets.

The coefficients of our final model are shown below:

```
print(summary(TV_Radio_model)$coefficients)
```

```
##               Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 2.92109991 0.294489678  9.919193 4.565557e-19
## TV          0.04575482 0.001390356 32.908708 5.436980e-82
## Radio       0.18799423 0.008039973 23.382446 9.776972e-59
```

With extremely low p-values, it is clear that there is a relationship between both varieties of advertisements and Sales. We can now make predictive judgements based off of our estimates and confidence intervals. For instance, in a given city with $100000 spent on TV advertisements and $20000 spent on radio advertisements, we can estimate with 95% confidence that sales should fall between 7,930 units and 14,580 units.

Conclusions

Based on this analysis, it is clear that TV and Radio advertisements both have a positive correlation with product sales. As TV advertising budget increases, sales tend to increase by (on average), 46 per $1000 of TV advertising. As Radio advertising budget increases, sales tend to increase by (on average), 188 per $1000 of Radio advertising.

The implications of this for both advertisers and media companies will depend on circumstances, but in general, our results support the hypothesis that TV and Radio advertisements are effective in increasing product sales. In addition, our predictive model can be used to determine how to allocate advertising budgets for products - based on our results here, it seems that radio advertisements tend to be most effective, but nonlinear factors such as diminishing returns on advertisements might cause an advertisers to choose to use television ads as well. In either case, our results indicate that newspaper ads are ineffective and should be avoided.