Stat 159 HW03

Abigail Chaver

October 14, 2016

Abstract

In this paper, we analyze a dataset containing continuous values for `Sales`, `Newspaper`, `Radio`, and `TV`. We attempt to establish a linear relationship regressing `Radio`, `Newspaper` and `TV` on `Sales` through, providing some visualization of the data and results of the model. This paper is supported by a similar analysis in Chapter 3 of *An Introduction to Statistical Learning*, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

Introduction

We are typically concerned with two types of analysis: prediction and inference. In the case of prediction, we are interested in converting a vector of predictors into a target variable through some model: our primary concern is to accurately predict what the target variable will be given predictor values that we haven't seen before. In inference, we are more interested in providing interpretable results about the relationship between two or more quantities, where we may hope that there is a causal relationship.

In this analysis, we are focused on inference: we wish to understand how spending on different types of media advertising impacts sales. If there is a strong positive relationship between Sales and one of the media types, it may benefit the bottom line to increase or maintain spending in that area; if there is no relationship or possibly a negative correlation, we may consider decreasing the budget.

Data

We are analyzing a set of 200 observations across 4 variables: the target, `Sales`, and three predictors, `TV`, `Radio`, and `Newspaper`. `Sales` represents number of sales of a particular product in thousands, while `TV`, `Radio`, and `Newspaper` represent dollar amounts of the advertising budget for each form of media respectively.

We can do a preliminary visual exploration of the variables to get a sense of their distributions - see page 2 and 3.

We then proceed to modeling the relationship between the variables.

Methodology

A linear model is almost never *true*: if there is a relationship between two variables, it is almost invariably more complicated than linear. However, linear models are surprisingly useful for prediction, and are very easy to interpret. They're also quite easy to compute - these three properties have made them popular in analysis for a long time.

A multiple variable linear model (in this case, 3 variables without interactions) is often written as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X2 + \beta_3 X3$$

Where Y is the target variable and the X_i are the predictor or input variables.

For this analysis, our model is

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper$$

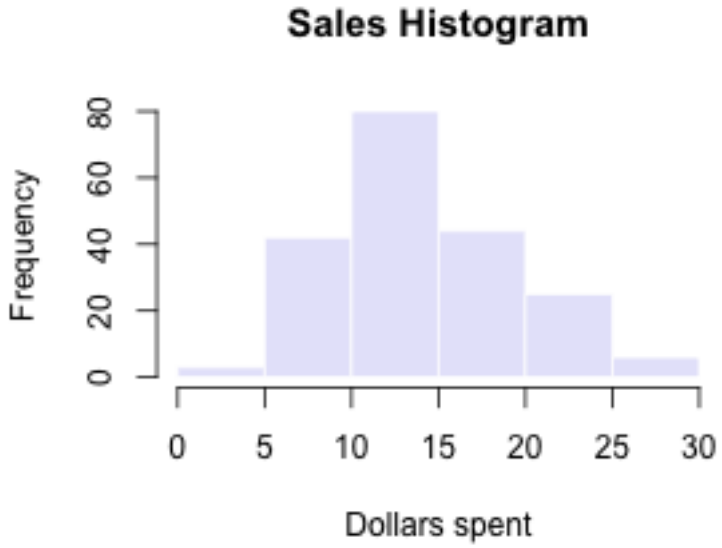

Figure 1: The distribution of sales across markets looks approximately normal.

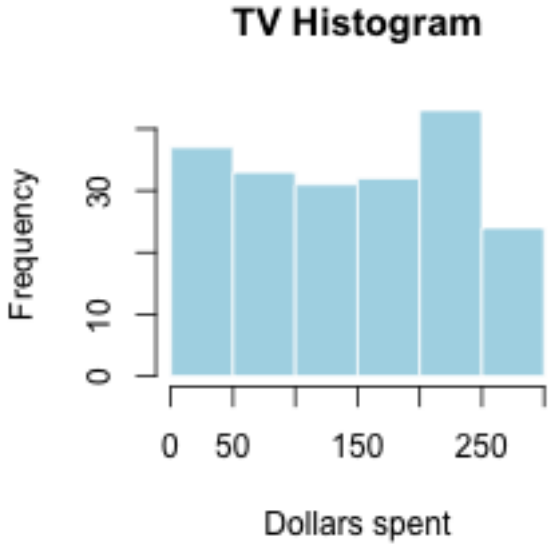

Figure 2: Spending on TV advertising looks fairly uniform.

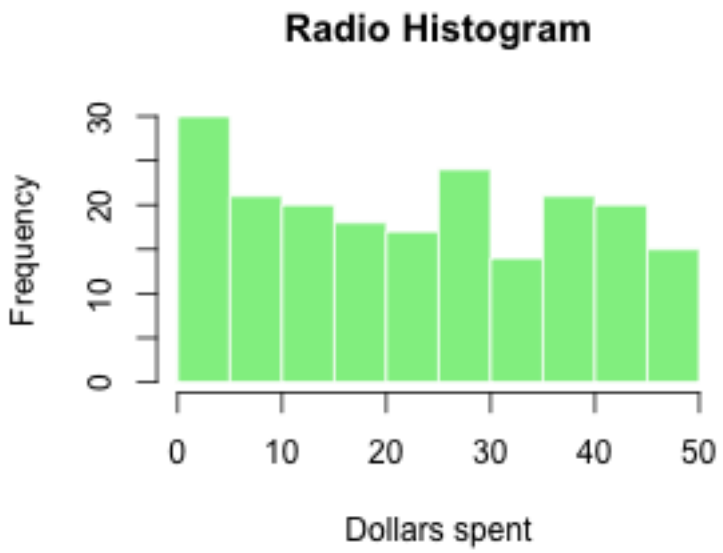

Figure 3: Spending on Radio advertising is skewed lower.

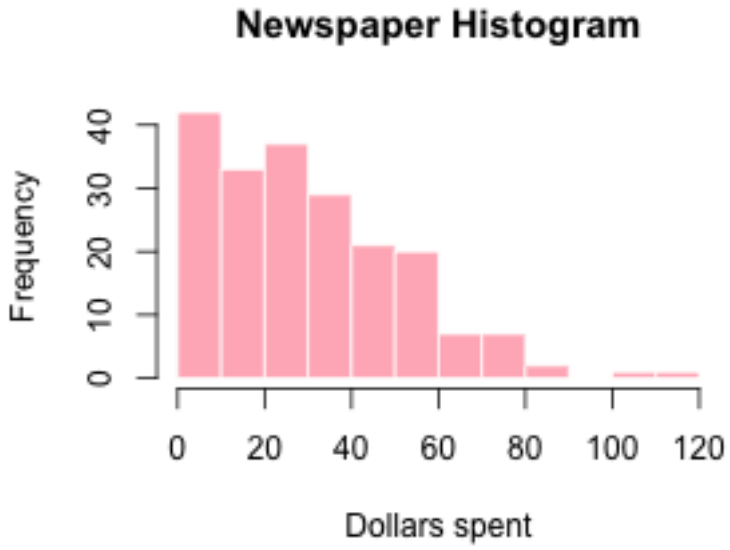

Figure 4: Spending on Newspaper advertising skews extremeley low.

$\hat{\beta}_0$ is simply an intercept which translates the line of best fit. The $\hat{\beta}_i$ are the values we are particularly interested in. A positive $\hat{\beta}_i$ indicates a positive correlation: more of X_i spending is associated with higher Sales. We are interested in how large $\hat{\beta}_i$ might be, as this indicates the strength of the relation. However, we must also look at its *significance* to understand whether a relationship actually exists - just because we find a nonzero $\hat{\beta}_i$ does not imply that there is a true relationship. We use the t-test to evaluate how likely it is that the true β_i is actually zero, and that our estimate $\hat{\beta}_i$'s distance from zero is due to chance.

Results

First we do some exploration: we can perform a scatterplot to get a visual impression of their relationships, and we will plot our linear model on top to see how accurate our model seems.

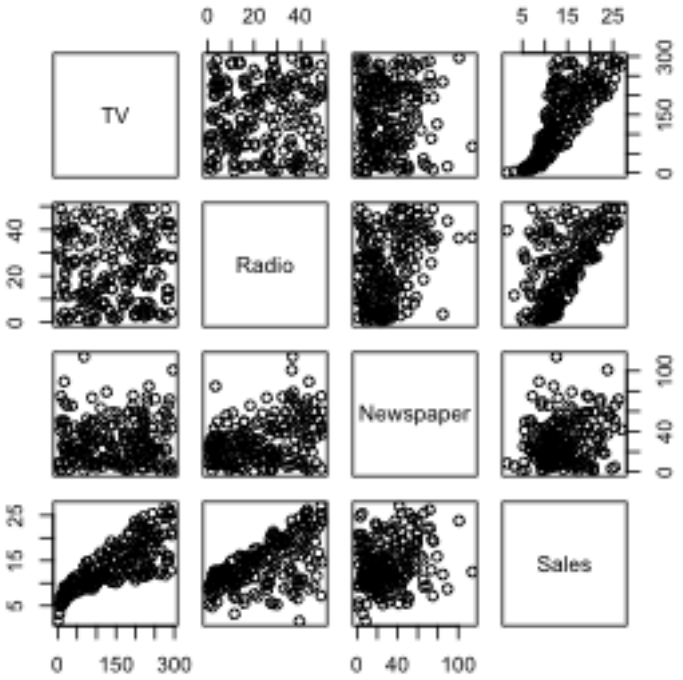

Figure 5: Scatterplot matrix of all variables, with regression line

This matrix can help us get a sense of whether a relationship exists between predictor and target, and can help us identify collinearity between variables, which can create problems in linear regression. We can also look at the quantitative correlations to help us.

Table 1: Correlation Matrix between all variable

	TV	Radio	Newspaper	Sales
TV	1	0.05481	0.05665	0.7822
Radio	0.05481	1	0.3541	0.5762
Newspaper	0.05665	0.3541	1	0.2283
Sales	0.7822	0.5762	0.2283	1

Fortunately, the relationship between `TV` and `Radio` seems very random- there does not seem to be any relationship. But `Neswpaper` seems somewhat correlated with `Radio`.

`TV` and `Radio` seem to have strong relationships with Sales. We take a closer look at the relationship between Sales and the three target variables, this time including the regression lines for one-variable models.

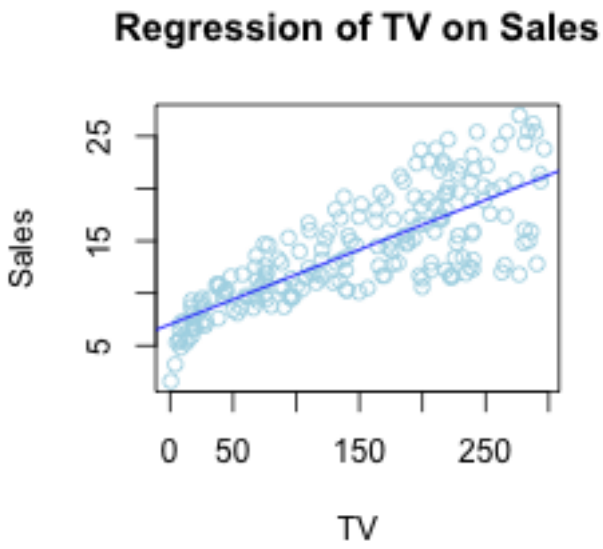

Figure 6: We can see that the relationship is positive, but clearly non-linear. The data is heteroskedastic and looks more like a log function than a linear one. However, our linear model may still be useful for interpretation.

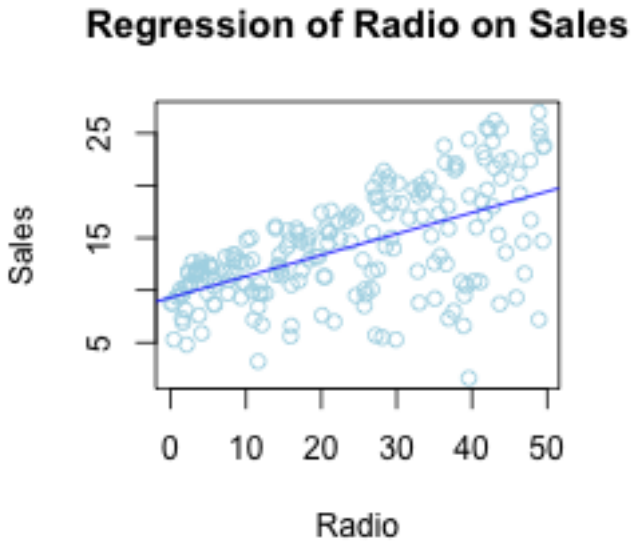

Figure 7: This relationship is also positive but seems nonlinear, and the spread of the data is larger.

Figure 8: Here there does not seem to be much of of a relationship. Newspaper seems unlikely to be a useful predictor in our model.

Then we compute the linear model with `lm(Sales ~ .)` to perform inference:

Table 2: Summary of Linear Regression Model of TV on Sales

	Estimate	Std. Error	t value	Pr(>|t|)
TV	0.04576	0.001395	32.81	1.51e-81
Radio	0.1885	0.008611	21.89	1.505e-54
Newspaper	-0.001037	0.005871	-0.1767	0.8599
(Intercept)	2.939	0.3119	9.422	1.267e-17

We first consider β_1 and its signficance. While the coefficient seems small, at `0.05`, we see that the t-statistic is quite large, making it very significant. In comparison, β_2 is larger, but its t-value is smaller (although still quite signficant). It's important to keep in mind that the value of a coefficient is highly dependent on the units of the variables and how they are scaled, so we are most interested in whether the β_i's are positive and how significant they are. We see that β_3, representing the linear relationship between Newspaper and Sales, is slightly negative but quite insignificant at a p-value of `0.86`. It seems that we may want to exclude Newspaper as a predictor for Sales, as it seems generally uncorrelated with sales, and somewhat correlated with TV and Radio.

We compute some other goodness-of-fit statistics to evaluate how far off our model is.

Table 3: Goodness of Fit Statistics

Statistic	Value
Residual Sum of Squares	556.8
R-squared	0.8972
Total Sum of Squares	5417
F-statistic	570.3
Residual Standard Error	1.686

We will also plot some visuals that can help us understand our model's fit.

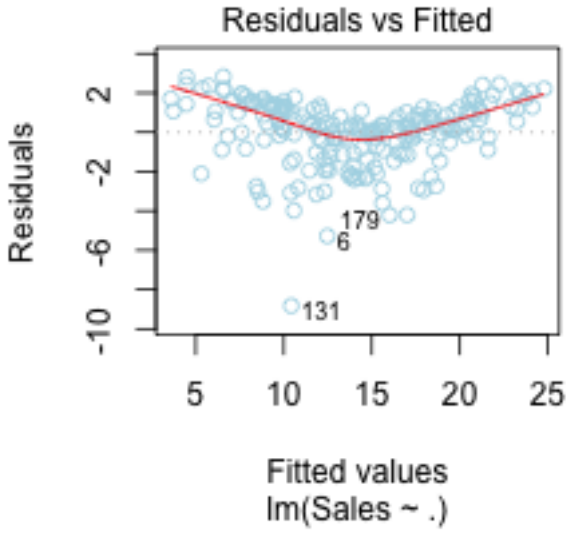

Figure 9: The residual plot confirms that the true relationship is not linear- when a linear model is accurate, there should be no trend in the residuals. Here there is a clear parabolic pattern indicating a nonlinear relationship between the predictors and target.

Conclusions

From our analysis, we can see that our model fits the data fairly well, even if the true relationship is not linear. Our R-squared value is fairly good, and this linear model may provide some value. If we wished to further explore the data, we would probably exclude Newspaper as a predictor and perhaps evaluate whether there is any significant interation between the two remaining predictors, TV and Radio.

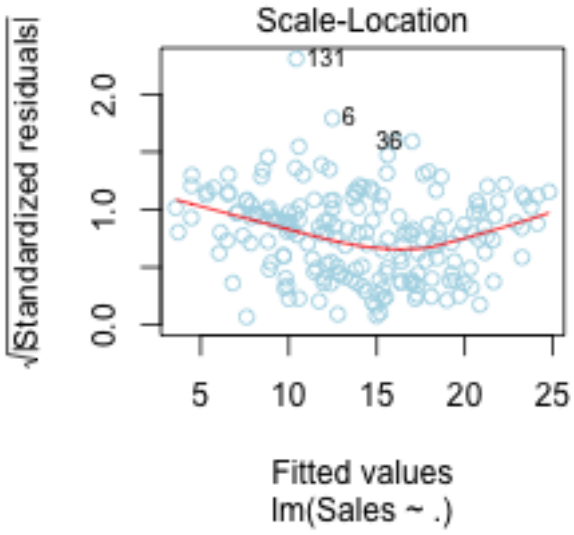

Figure 10: A Scale-Location plot is helpful for identifying whether data is heteroskedastic. Surprisingly, the spread seems fairly random and unrelated to the fitted values. There is a slight trend, but generally heteroskedasiticity does not seem to be a large problem in our model.

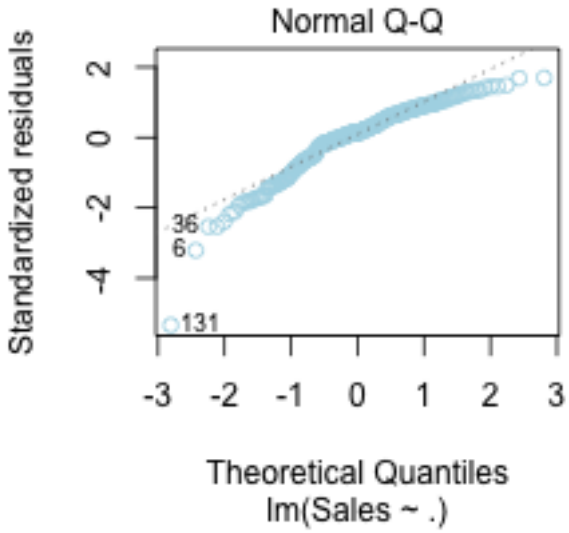

Figure 11: Here we evaluate whether the data seems normally distributed, which is an assumption in a linear regression model. The data is not perfectly normal, especially in the tails. However, the data is reasonably close to a normal distribution, and this may explain why our goodness-of-fit statistics are fairly strong.