Stat 159 HW02

Abigail Chaver

October 7, 2016

`{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)`

Abstract

In this paper, we analyze a dataset containing continuous values for `Sales`, `Audio`, `Radio`, and `TV`. We attempt to establish a linear relationship between `Sales` and `TV` through simple one-variable linear regression, 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 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 television advertising impacts sales. If there is a strong positive relationship, it may benefit the bottom line to increase or maintain TV spending; if there is no relationship or possibly a negative correlation, we may consider decreasing TV budgets.

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. In this analysis we will only consider `TV`, of the three predictors.

We can do a preliminary visual exploration of the two relevant variables, `TV` and 'Sales', to get a sense of their distributions (See Figures 1 and 2 on Page 2).

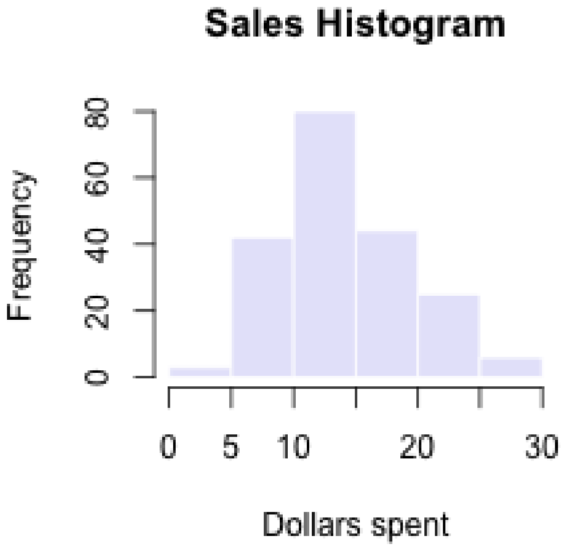

Figure 1: The distribution of sales across markets looks approximately normal.

We now proceed to modeling the relationship between the two 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 one variable linear model is often written as

$$Y = \beta_0 + \beta_1 X$$

Where Y is the target variable and X is the predictor or input variable.

For this analysis, our model is

$$Sales = \beta_0 + \beta_1 TV$$

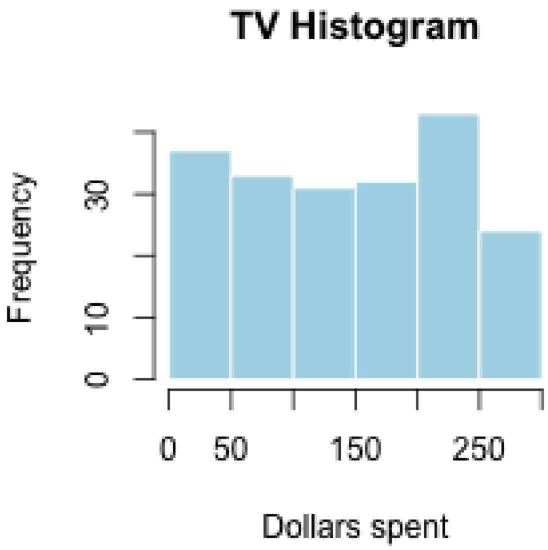

Figure 2: Spending on TV advertising looks fairly uniform.

β_0 is simply an intercept which translates the line of best fit. β_1 is the figure we are particularly interested in. A positive β_1 indicates a positive correlation: more TV spending is associated with higher Sales. We are interested in how large β_1 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 β_1 does not imply that there is a true relationship. We use the t-test to evaluate how likely it is that the true β_1 is actually zero, and that our estimate $\hat{\beta}_1$'s distance from zero is due to chance.

Results

We compute the linear model with `lm(Sales ~ TV)` to perform inference:

`{r, echo=FALSE} load("data/regression.Rdata") library(pander) panderOptions("digits", 2) pander(fit, caption = "Summary of Linear Regression Model of TV on Sales")`

As mentioned before, we are particularly interested in 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. 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 β_1 is positive and how significant it is. In this case, there is a strong positive correlation between TV budget and sales figures.

We will now evaluate how well the model fits the data. We can perform a scatterplot to get a visual impression of their relationship, and we will plot our linear model on top to see how accurate our model seems.

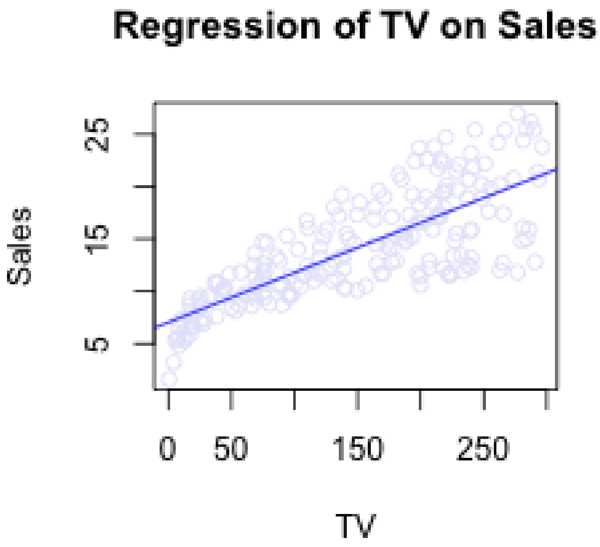

Figure 3: Scatterplot of TV and Sales, with regression line

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

```
{r, echo=FALSE} Rsq <- summ$r.squared RSE <- summ[[6]] fstat <-
summ$fstatistic['value'] Statistic <- c("R-squared", "Residual
Etandard Error", "F-statistic") Value <- c(Rsq, RSE, fstat)
fit_table <- data.frame(Statistic, Value) pander(fit_table,
caption="Goodness of Fit Statistics")
```

Conclusions

From our analysis, we can see that our model fits the data acceptably but not *exceptionally*. There is a strong positive relationship between our predictor and target variable, but upon visual inspection we can see that the relationship is not linear. However, our R-squared value is not atrocious, and this linear model may provide some value.