Multiple Linear Regression Analysis

Bryana Gutierrez

October 14, 2016

Abstract

The analysis is an attempt to reproduce the results found in Section 3.2 of *Multiple Linear Regression* (chapter 3) of the book **An Introduction to Statistical Learning**. This is an exploration of Multiple Linear Regression.

Introduction

This analysis takes Advertising data and attempts to map a linear relationship between various advertising budget (TV, radio, and newspaper) and product sales. The best way to do this is through the method of least squares.

Data

In this analysis we take data from 200 distinct markets. This data is contained in `Advertising.csv` which has five variables: `X` a counter, `Sales` the product sales in thousands of units, and `TV`, `Radio`, and `Newspaper` the advertising budgets for each medium in thousands of dollars. In this multiple linear regression case, we look at how all three advertising budgets, `TV`, `Radio`, and `Newspaper`, correlate to `Sales`.

Methodology

As the title of this report suggests, this is a multiple linear regression analysis. We use the linear model

$$y \approx \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3$$

to describe the relationship between `Sales`, `TV`, `Radio`, and `Newspaper`. Therefore, the linear model looks more like this:

$$Sales \approx \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper$$

where β_0 is the intercept term and the $\beta_i s$ describe how each advertising budget affects the sales. As mentioned before, the best way to estimate the variables in this model is through the least squares method.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

is the least squares estimate of β which contains the actual values of the $\beta_i s$. By the Gauss-Markov Theorem they are the best linear unbiased estimators. They are estimated by minimizing the sum of the residual squared errors (RSS):

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

where e_i is equal to $y_i - \hat{y}_i$. $\hat{y}_i$ is calculated by using the model and $\hat{\beta}$:

$$\hat{y}_i = X\hat{\beta}$$

where

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ 1 & x_{2,1} & x_{2,2} & x_{2,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{200,1} & x_{200,2} & x_{200,3} \end{bmatrix}$$

$\hat{y}_i$ is the predicted y value. In terms of this analysis, $\hat{y}_i$ is the amount of predicted sales based off of the all the different advertising budgets.Basically, minimizing the RSS would be minimizing the error of the prediction.

RSS can also be written as:

$$RSS = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 * x_{i,1} - \hat{\beta}_2 * x_{i,2} - \hat{\beta}_3 * x_{i,3})$$

Minimizing this value over the $\hat{\beta}_i s$ results in

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where Y is a vector with all the y values. Using the `Advertising.csv` data we replace the $y_i s$ with the `Sales` numbers, the $x_{i,1}$ with the `TV` numbers, the $x_{i,2}$ with the `Radio` numbers, and the $x_{i,3}$ with the `Newspaper` numbers.

Results

Using R we find the values for $\hat{\beta}$ and information about their accuracy. First, we look at how each advertising budget affects `Sales` on its own.

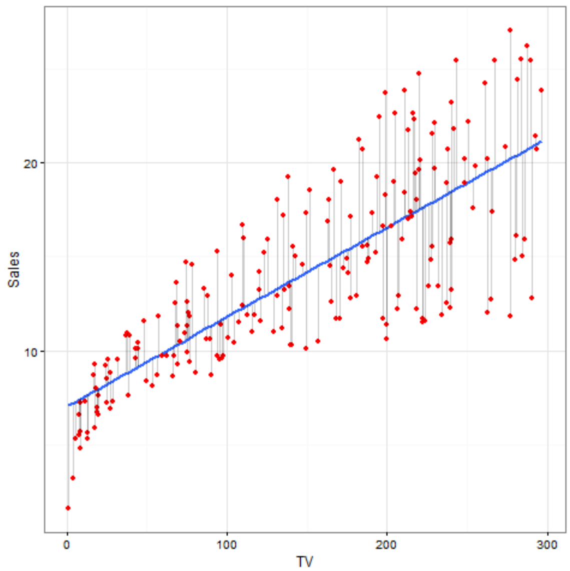
TV vs. Sales

Looking at the `TV` and `Sales` data, we perform a simple linear regression. $Sales \approx \beta_0 + \beta_1 * TV$. **Table 1** is the output for what R calculates for the estimates of these β values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.0326	0.4578	15.36	0.0000
TV	0.0475	0.0027	17.67	0.0000

Table 1: Information About Regression Coefficients in a Reduced Model

This the scatter plot with the best fit regression line that uses the coefficients above.



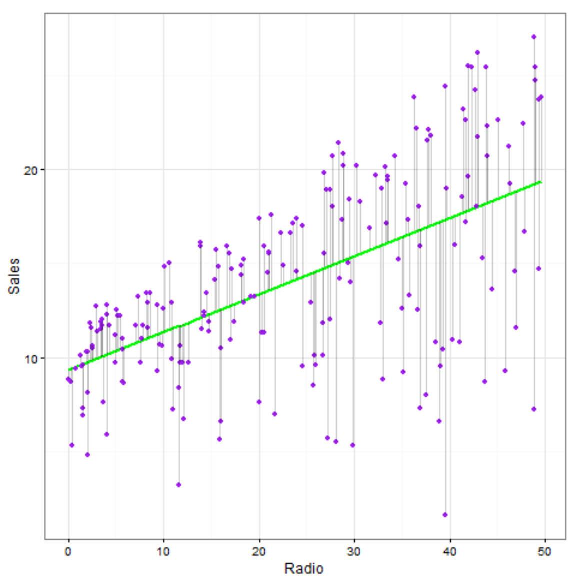
Radio vs. Sales

Looking at the `Radio` and `Sales` data, we perform a simple linear regression. $Sales \approx \beta_0 + \beta_1 * Radio$. **Table 2** is the output for what R calculates for the estimates of these β values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.3116	0.5629	16.54	0.0000
Radio	0.2025	0.0204	9.92	0.0000

Table 2: Information About Regression Coefficients in a Reduced Model

This the scatter plot with the best fit regression line that uses the coefficients above.



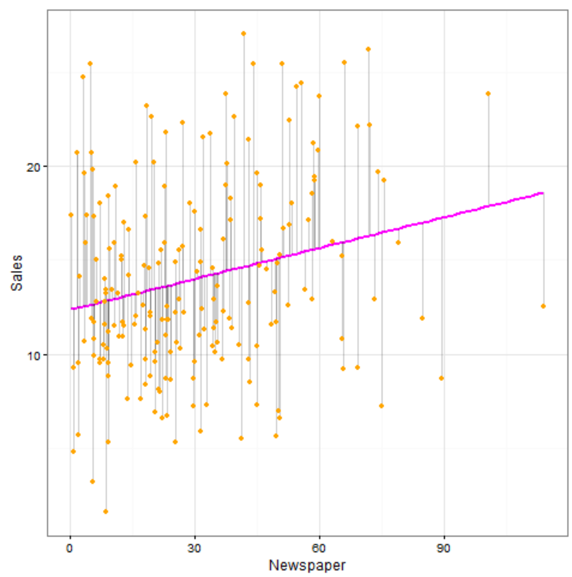
Newspaper vs. Sales

Looking at the `Newspaper` and `Sales` data, we perform a simple linear regression. $Sales \approx \beta_0 + \beta_1 * Newspaper$. **Table 3** is the output for what R calculates for the estimates of these β values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.3514	0.6214	19.88	0.0000
Newspaper	0.0547	0.0166	3.30	0.0011

Table 3: Information About Regression Coefficients in a Reduced Model

This the scatter plot with the best fit regression line that uses the coefficients above.



Full Model

Now we look at the entire multiple linear model. **Table 4** is a table of the $\hat{\beta}_i$ that we found using the method described earlier: $(X^T X)^{-1} X^T Y$. This table also includes information about the accuracy of these estimates.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9389	0.3119	9.42	0.0000
TV	0.0458	0.0014	32.81	0.0000
Radio	0.1885	0.0086	21.89	0.0000
Newspaper	-0.0010	0.0059	-0.18	0.8599

Table 4: Information About Regression Coefficients in the Full Model

`Std. Error` is a measure of the volatility of the estimates and the last two columns are indicators of the validity of the estimate. In this case since the p-values (the last column) are practically zero for all the $\hat{\beta}_i s$ expect for newspaper. This indicates that the estimates for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are validly nonzero. However, the estimate for $\hat{\beta}_4$ has more of a change of being zero and not affecting the model. From this we can tell that `TV` and `Radio` are better predictors of `Sales` than `Newspaper` is.

Analyzing The Estimates

The following statistics validate the accuracy of the linear model

$$Sales \approx \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper$$

One example of such a measure is the correlation matrix. This matrix contains the covariances of each $\hat{\beta}_i$ with every other $\hat{\beta}_i$. Note that the covariance of $\hat{\beta}_i$ with itself is just the variance of $\hat{\beta}_i$. The covariance matrix can be seen in **Table 5**

	X	TV	Radio	Newspaper	Sales
X	1.00	0.02	-0.11	-0.15	-0.05
TV		1.00	0.05	0.06	0.78
Radio			1.00	0.35	0.58
Newspaper				1.00	0.23
Sales					1.00

Table 5: Covariance Matrix

Analyzing the Model

More examples of measures of accuracy of the model are the RSE (residual standard error), R^2 statistic, and $F - Statistic$. **Table 6** shows these values.

Quantity	Value
RSE	1.69
R2	0.90
F-Stat	570.27

Table 6: Regression Quality Statistics

RSE

`RSE` is the residual standard error, which is a measure of the accuracy of the predicted values of `Sales` that you can get from the model. In mathematical terms, this is

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

This adds the differences between the actual and predicted values of the y's which in this case is the `Sales` numbers. `RSE` equal to 1.6855104 indicates that the predicted `Sales` number is off by approximately 1685.5103734 units.

R Squared

The R^2 statistic measures proportionally how much of the variability of `Sales` can be due to `TV`, `Radio`, and `Newspaper`. Mathematically,

$$R^2 = \frac{TSS - RSS}{RSS} = 1 - \frac{RSS}{TSS}$$

where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$. The closer the R^2 statistic is to one, the better the multiple linear model is at modeling `Sales`. In this case the R^2 statistic is 0.8972106, so it is pretty close to one.

F-Statistic

The F-Statistic is a measure of how good the model is. It uses the `RSS` and the `TSS` just like the R^2 statistic, but also incorporates the F distribution. Mathematically, the F-statistic is

$$F - Stat = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

The p-value for this F-statistic is $1.5752273 \times 10^{-96}$ which is much smaller than any of the p-values for each of the $\hat{\beta}_i s$. This is also much smaller than the p-values computed in the simple linear regression that assess the variables `TV`, `Radio`, and `Newspaper` for their ability to predict the response variable, `Sales`. These p-values are $1.4673897 \times 10^{-42}$ for `TV`, 4.354966×10^{-19} for `Radio`, and 0.0011482 for `Newspaper`. Therefore, although individually, the variables are decent predictors of `Sales`, collectively, they are better.

Conclusion

Using the statistics on the $\hat{\beta}_i s$, the model analysis, and the visual representations of the least squares fitted line, we can see that the model $Sales \approx \beta_0 + \beta_1 * TV + \beta_2 * Radio + \beta_3 * Newspaper$ was a reasonable assumption.