Linear Regression Report

Jacqueline Liu

October 15, 2016

Abstract:

Linear regression is a simple, yet powerful method of modeling and supervised learning. Many real-world phenomena can be represented as a linear model, or with other models that employ linear regression. The well-known and well-referenced text **An Introduction to Statistical Learning** covers this technique in chapter 3, sections 1 and 2, which cover simple and multiple linear regression. Here we will reproduce their results to highlight key steps and ideas.

Introduction:

There is a wide variety of applications for linear regression; it can be used for prediction of response to a medical treatment, forecasting prices of fruits a month from now, or determining whether or not sleep and grades are correlated. Ultimately, the goal is to describe some trend or relationship in a dataset. Once the appropriate model is found, predictions can be made for certain variables if given the values of others and the strength of relationships can be assessed. Here, we'll be looking at a commercial setting and determining the relationship between advertising spending and sales of a given product. Specifically, we want to predict sales given the advertising budget for three different media outlets.

Data:

The `Advertising` dataset was 'collected' from 200 different markets. It records the advertising budgets (in thousands of dollars) for `TV`, `Radio`, and `Newspaper` as well as the number of `Sales` (in thounds of units) for each of the markets. Because `Sales` is what we're ultimately trying to predict, it is our response variable while the other three columns will be our features.

Methodology:

We're using linear models, which can be written as:

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

Given X and Y, the coefficients β_0 and β_1 are chosen to minimize the sum of squared errors:

$$SSE = \sum (Y - \hat{Y})^2$$

giving our regression line the name "least squares regression". There are other loss metrics we could use (weighted errors, least absolute deviation, L1-norm penalty) but least squares is one of the most widedly used. We can regress `Sales` on `TV`, then on `Radio`, and `Newspaper` to find three different linear models. We will also regress `Sales` on all three variables at once to better understand their high-dimensional relationship, following the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

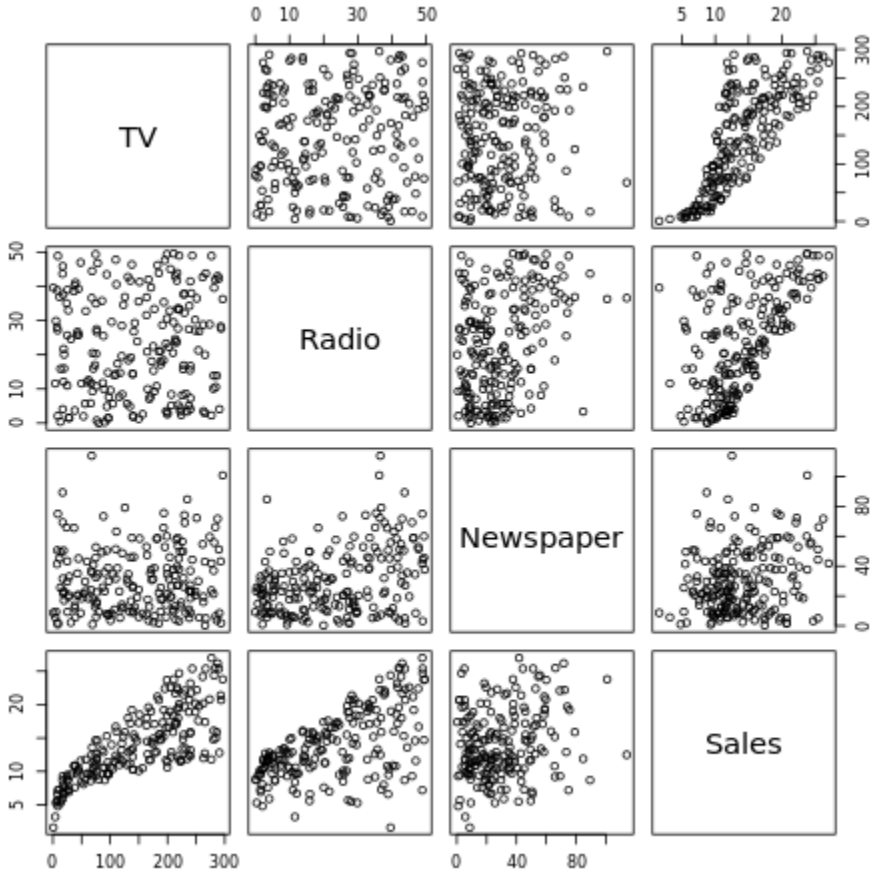
Results:

To understand the linear relationship between the features, here is the correlation matrix:

	TV	Radio	Newspaper	Sales
TV	1.00	0.05	0.06	0.78
Radio	0.05	1.00	0.35	0.58
Newspaper	0.06	0.35	1.00	0.23
Sales	0.78	0.58	0.23	1.00

Table 1: Correlation Matrix

and here is the scatterplot matrix:



The regression coefficients for the simple linear regression models are:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	7.03	0.46	15.36	0.00
TV	0.05	0.00	17.67	0.00

Table 2: Sales onto TV

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.31	0.56	16.54	0.00
Radio	0.20	0.02	9.92	0.00

Table 3: Sales onto Radio

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12.35	0.62	19.88	0.00
Newspaper	0.05	0.02	3.30	0.00

Table 4: Sales onto Newspaper

and their quality indices:

	Quantity	Value
1	RSS	3.26
2	R2	0.61
3	F-stat	312.14

Table 5: TV Regression Quality Indices

	Quantity	Value
1	RSS	4.27
2	R2	0.33
3	F-stat	98.42

Table 6: Radio Regression Quality Indices

	Quantity	Value
1	RSS	5.09
2	R2	0.05
3	F-stat	10.89

Table 7: Newspaper Regression Quality Indices

(NOTE: the p-values are not exactly 0, but so small that for all intents and purposes they basically are). For the multiple linear regression, the coefficients are:

The multiple regression model gives rise to the following coefficients and quality indices:

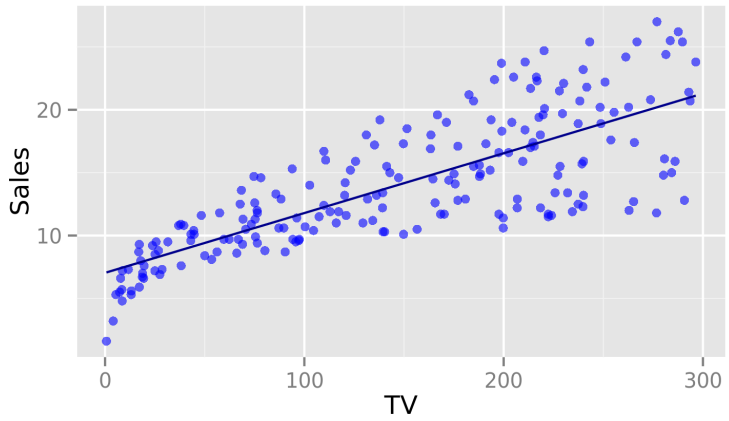
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.94	0.31	9.42	0.00
TV	0.05	0.00	32.81	0.00
Radio	0.19	0.01	21.89	0.00
Newspaper	-0.00	0.01	-0.18	0.86

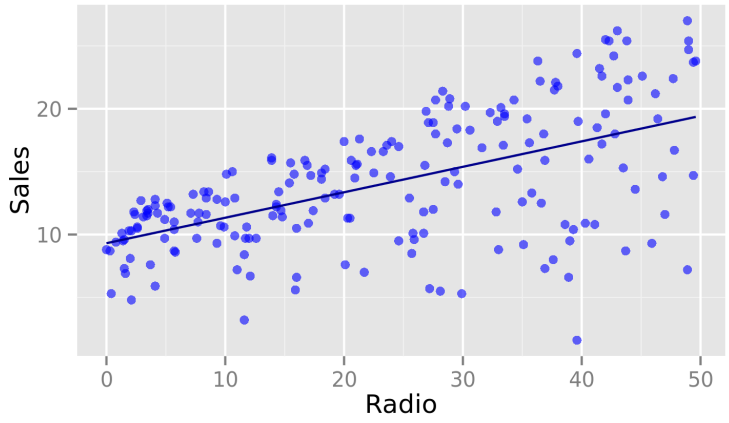
Table 8: Sales onto TV, Radio, and News

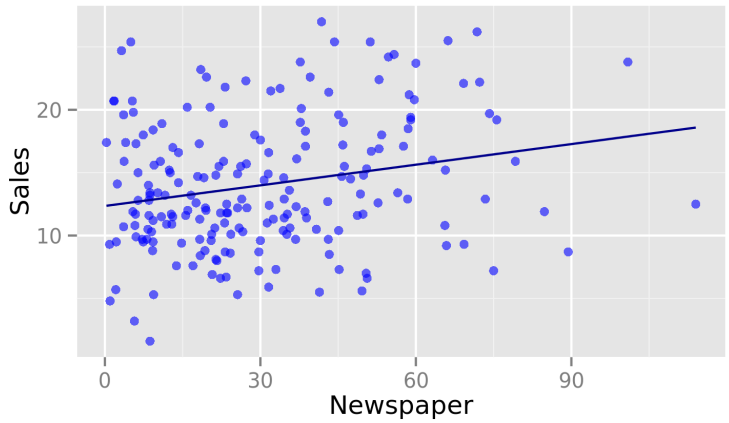
	Quantity	Value
1	RSS	1.69
2	R2	0.90
3	F-stat	570.27

Table 9: Multiple Regression Quality Indices

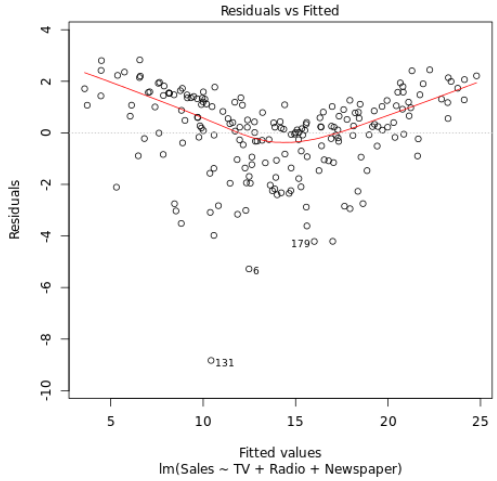
To visualize the data and the proposed simple linear models, here are scatterplots with the calculated regression lines:



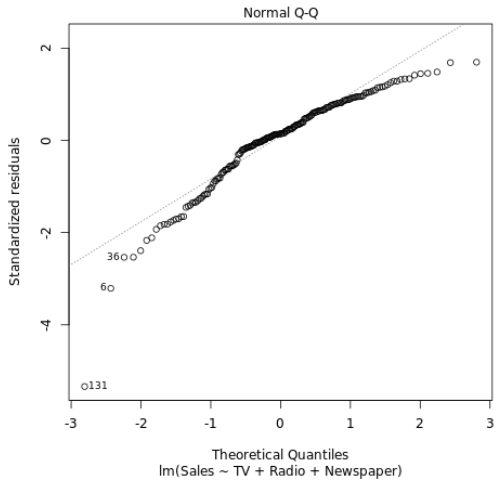




The features collectively produce a model with the following residual plot:



and normal qq plot:



Conclusion:

Looking at the results for the simple linear models, we can see though the intercepts and slopes are significantly different from 0 (meaning there is a relationship between the advertising budges and 0). However the slopes themselves are not all very large; only `Radio` has a slope in the tenths, vs hundreths, decimal place. However, this isn't to say that the company should spent all its advertising budget on radio advertisements. `TV` has the greatest R^2-value, meaning it is best at explaining the variation in `Sales`, which can be clearly seen in the scatterplots.

However, if we look at the multiple regression model, `Newspaper` is no longer has any significant impact on sales; its slope is now nearly 0 and p-value 0.86. This could be explained by the 0.35 correlation between `Newspaper` and `Radio`; markets with higher spending on news ads are spending more on radio ads too, and it's the radio ads that are boosting sales, not the news. `Radio` and `TV` have regression coefficients that don't vary much between the simple and multiple regression models, and together produce an R^2 value of 0.90, which is almost 0.30 better than either had individually. This demonstrates the importance of considering the influence of the relationships between features on our predictors, rather than trying to isolate each. There are also non-linear relationships to consider; the normal qq plot shows a bit a skew so our model is biased. The more hyperparameters involved, however, the more likely the model is prone to overfitting. Hence model selection and hyperparameter tuning always involves a bias-variance tradeoff.