3.1 Simple Linear Regression

Zongyi Liu

2023-05-13

In previous `advertising` and `sales` dataset, we might want answer those questions:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

3.1 Simple Linear Regression

Simple Linear Regression can be expressed as

$$Y \approx \beta_0 + \beta_1 X.$$

This is said as **regressing Y on X**.

Here, y-hat indicates a prediction of Y on the basis of X=x, and betas are called **coefficients** or **parameters**.

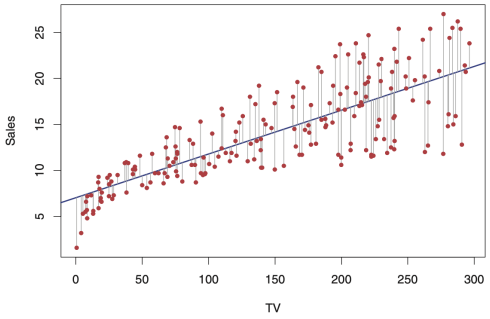
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

3.1.1 Estimating the Coefficients

We want to know the value of β_0 and β_1, and we have observations:

$$(x_1, y_1),\ (x_2, y_2), \ldots,\ (x_n, y_n)$$

And we know that $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for i=1,... n. The rationale to get the approximations for parameters is by the **least squares criterion**.



Then we have $e_i = y_i - \hat{y}_i$, and we have the **residual sum of squares (RSS)** defined as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

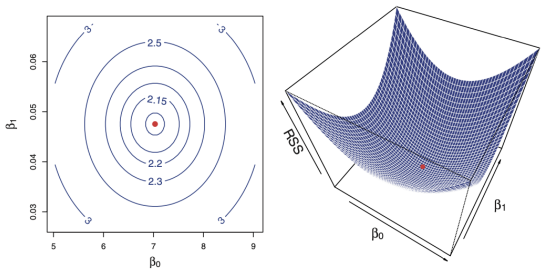
Then we will have

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where

$$\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i \text{ and } \bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i \text{ are the sample means}$$

We can illustrated this as

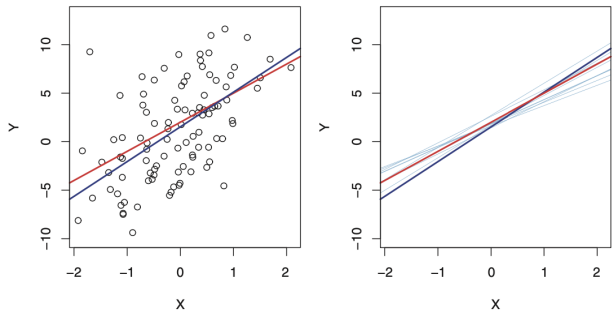


3.1.2 Assessing the Accuracy of the Coefficient Estimates

We know that in the population there is a function between Y and X which has the following relationship:

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

There are two lines in our definition: population regression line, which is true, and least squares line, which is what we got. In this plot, the red line is the population regression line, and the blue line is the least squares line.



To know how close the $\hat{\beta}_0$ and $\hat{\beta}_1$ is to β_0 and β_1, we will introduce the concept of standard errors, which can be computed as follows:

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

Here $\sigma^2 = \mathrm{Var}(\epsilon)$., its squared value is known as the residual standard error, which is given by $RSE = \sqrt{RSS/(n-2)}$, then we can introduce the concept of confidence interval, which is given by

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

and

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0).$$

They can be used to perform the hypothesis testing. There are two hypotheses, the null hypothesis is

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

and the alternative hypothesis is

$$H_a : \text{There is some relationship between } X \text{ and } Y.$$

This corresponds to testing

$$H_0 : \beta_1 = 0$$

and

$$H_a : \beta_1 \neq 0,$$

To verify the hypothesis, we will need to compute a t-statistics and its p-value, and compare it with the 1 or 5% p-value to determine whether to reject the null hypothesis or not.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

3.1.3 Assessing the Accuracy of the Model

There are two statistics for us to use to assess the accuracy of the model, residual standard error (RSE) and the R^2 statistic.

Residual Standard Error

It is computed as

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

and RSS is given as

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

If the RSE is small, then we can say that the model fits the data well.

R^2 statistic

It is computed as

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

R^2 can be literally interpreted as the proportion of variability in Y that can be explained using X.

An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

There is another measurement to test the accuracy of the model, which is correlation:

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

In simple regression, they are the same, but in multiple regress, they are not.