Chapter 3: Linear Regression

Esben Eickhardt

2023-05-23

Introduction

In this chapter we focus on *linear regressions*, how they work and what kind of problems they can solve.

What Questions Can We Answer?

Suppose we have a data set with sales as well as budgets for various types of advertisement, what kind of questions would we be able to answer:

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media are associated with sales?
- How large is the association between each medium and sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

1. Simple Linear Regression

Mathematically we can write a *simple linear regression* as:

$Y = \beta_0 + \beta_1 X$

This could e.g. be a relationship between a TV-advertisement budget and sales, where β_0 and β_1 are two unknown *coefficients* that represent the *intercept* with the y-axis and the *slope* of the line.

$Sales = \beta_0 + \beta_1 * TV$

1.1 Estimating the Coefficients

To estimate the coefficients we use data. That is, we use observations pairs of TV-advertisement and Sales. Our aim is to find the two *coefficients* that result in the lowest difference between the actual sales and the predicted sales. We calculate the distance between the actual sales and the predicted sales using *least squares*, which is just the squared difference. The difference is referred to as a *residual*. In **Figure 1** the *residuals* can be seen as grey lines.

A residual is calculated as the difference between the actual value and the predicted value:

$e = y_i - \hat{y}_i$

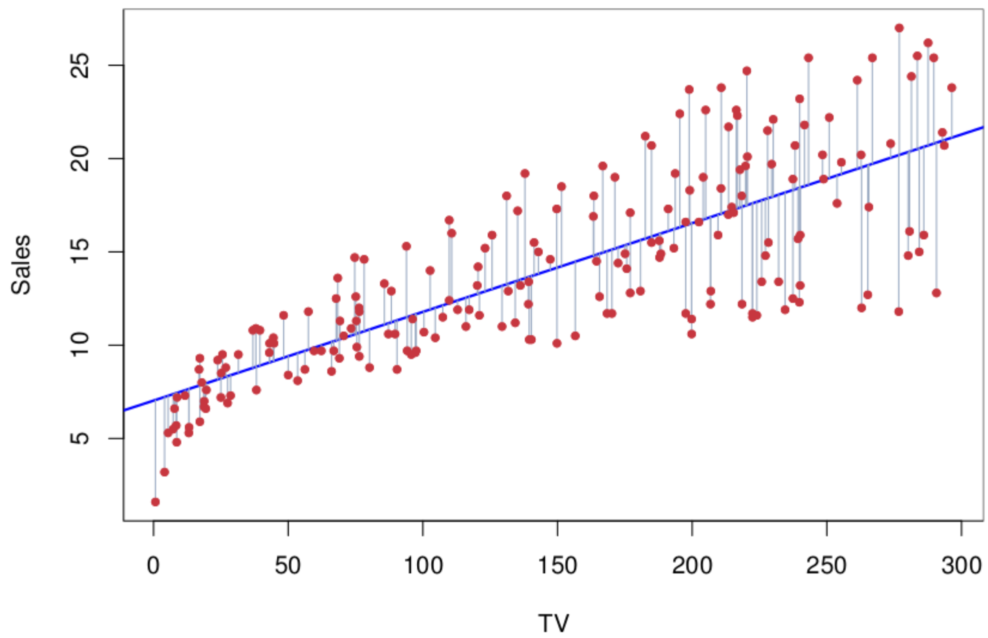

Figure 1: Linear Model for TV vs Sales

The *residual sum of squares (RSS)* is calculated as:

$RSS = e_1^2 + e_2^2 + ... + e_n^2 <=>$
$RSS = (y_1 + \beta_0 - \beta_1 x_1)^2 + (y_2 + \beta_0 - \beta_1 x_2)^2 .... + (y_n + \beta_0 - \beta_1 x_n)^2$

The two *coefficients*, β_0 and β_1, can be minimized with the following equations:

$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})}$
$\beta_0 = \overline{y} - \beta_1 \overline{x}$

In **Figure 2** you can see the calculated RSS for different β_0 and β_1 values used on the advertisement dataset. The red dot marks the minimum RSS.

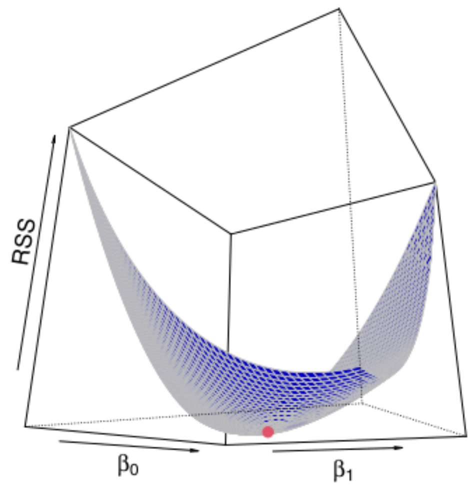

Figure 2: Minimizing RSS

1.2 Assessing the Accuracy of the Coefficient Estimates

We have a true model:

$Y = 2 + 3X + \epsilon$

We create 100 random Xs and calculate Ys where we add some random noise, ϵ. In **Figure 3** these noisy points are shown, the true *population regression line* is shown in red and the *least squares line* calculated on the noisy sample is shown in blue.

To the right in **Figure 3** we have made noisy samples ten times and calculated their regression lines (light blue).

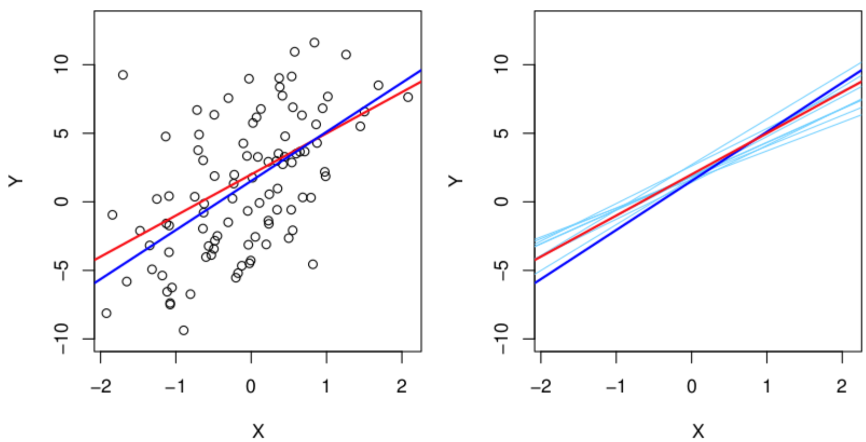

Figure 3: Regression Lines for Similated Data

In general the *sample coefficients* are good estimates of the *population coefficients*, and on average (μ) we expect them to be equal to the *population coefficients*. However, in a real world scenario we only have one sample, and we do not know if it is an over- or under-estimate of the *population coefficients*. In general our confidence in the *sample coefficients* is estimated by computing the *standard error*.

$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$

The *standard error* roughly tells us the average amount that the estimated sample $\hat{\mu}$ differs from the actual value of μ. It is calculated in the following way:

1. Collect a sample of data from the population you're interested in studying.
2. Calculate the mean of your sample
3. Calculate the difference between each observation in your sample and the sample mean $(x - \overline{x})$
4. Square each of the differences obtained in Step 3.
5. Sum up all the squared deviations obtained in Step 4.
6. Divide the sum of squared deviations from Step 5 by the sample size (n), which gives us the variance (the average squared deviation).
7. Calculate the standard error (SE) by taking the square root of the variance (s^2) and dividing it by the square root of the sample size (n).

The more observations we have the smaller the standard error of $\hat{\mu}$.

The Standard errors can be used to compute *confidence intervals*, e.g. a 95 confidence interval is a range of values such that with 95 % probability the range will contain the true value.

For linear regression the 95 % confidence interval for β_1 can be approximateted using:

$\hat{\beta}_1 \pm 2 * SE(\hat{\beta}_1)$

t-statistics is how the number of standard deviations a value is away from zero.

$t = \frac{\hat{\beta_1} - 0}{SE(\hat{\beta_1})}$

1.3 Assessing the Accuracy of the Model