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 total difference between the actual sales and the predicted sales. Despite this aim most models instead attempt to minimize the mean squared error (MSE), as it has better mathematical properties than the mean absolute error (MAE). The difference between a prediction and the actual value is referred to as the *residual*, and the residuals can be seen in **Figure 1** as grey lines.

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

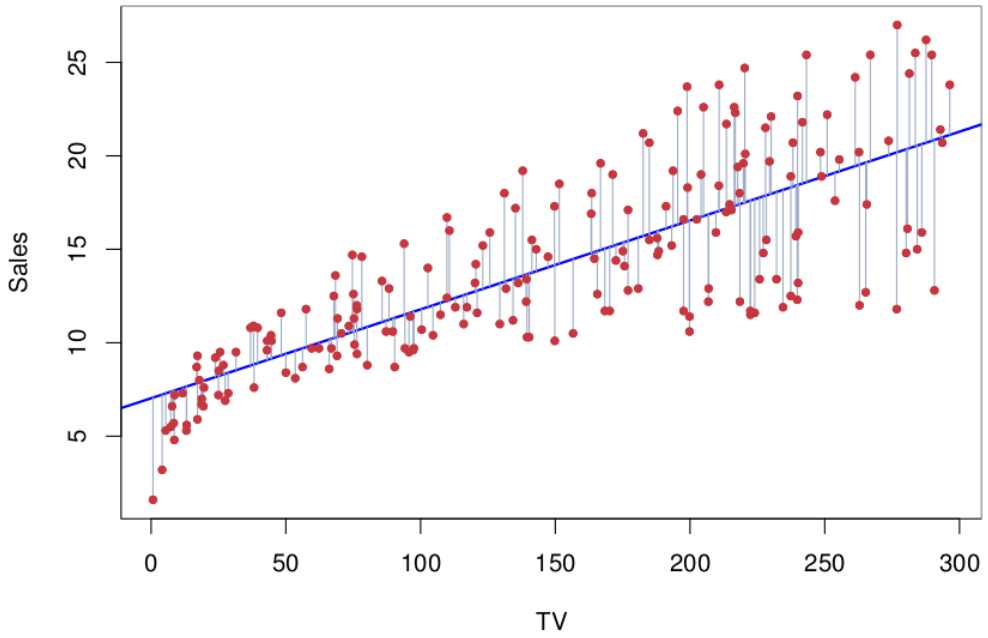

Figure 1: Linear Model for TV vs Sales

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

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.

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).

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* are estimated by computing the *standard error*. The *standard error* can be calculated via a formula (see below) or via *bootstrapping*. I prefer bootstrapping as it

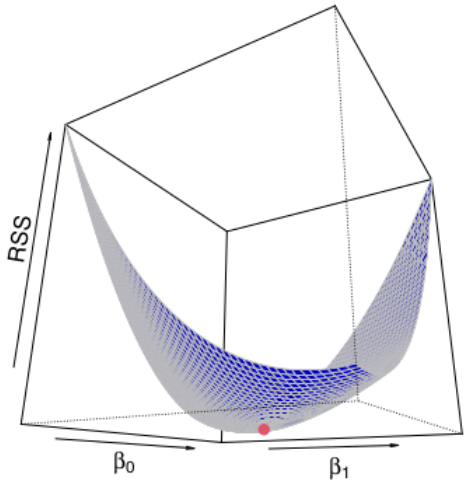

Figure 2: Minimizing RSS

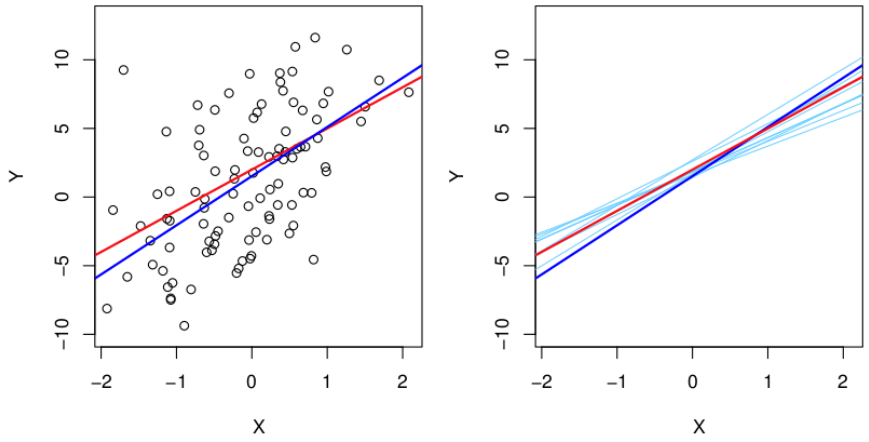

Figure 3: Regression Lines for Similated Data

is a more transparent solution, as you sample your sample, and identify the deviation in the sample mean (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 (sˆ2) 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*. A *95 % confidence interval* e.g. 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 approximated using:

$\hat{\beta_1} \pm 2 * SE(\hat{\beta_1})$

t-statistics is 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

Here we want to quantify *the extent to which the model fits the data*. This is usually assessed using the two quantities:

- Residual standard error (RSE)
- R^2 Statistics

The RSE is the average amount that the response will deviate from the true regression line. The formular uses a *magic number*, 2, to account for the error in small data sets.

$RSE = \sqrt{\frac{1}{n-2} RSS}$

We can calculate the percentage error by comparing the error if we had just predicted the mean value of y for all data points to the RSE. The *RSE/MeanError* tells us the percentage error.

R^2 provides an absolute measure of the lack of fit of the model, and is a number in the range from 0-1. It is a measure of the *proportion of variance explained*.

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

If a model fits the data perfectly *RSS* would be zero, and thus 100 % (R^2) of the variance would be explained by the model. TSS is the *total sum of squares*, which is $TSS = \sum (y_i - \overline{y})$. TSS is thus the *sum of squares* for a model that is simply a straight line/hyperplane through the mean of the data.

2. Multiple Linear Regression

I practice we have more than one predictor, thus here we extend the *linear regression* to multiple variables.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$

In such a model we interpret β_j as the average effect on Y of a one unit increase of X_j if all other predictors are fixed.

2.1 Estimating the Regression Coefficients

As with the simple linear regression we again want to minimize the RSS. See **Figure 4**.

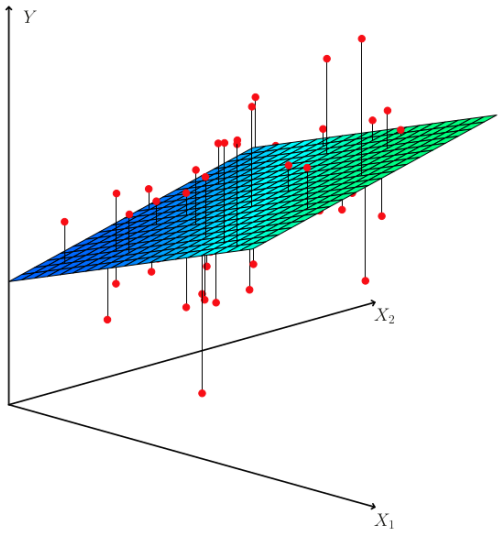

Figure 4: Multiple Linear Regression

The coefficients for a Multiple Linear Regression can be interpreted in the following way. For a given amount of X_1 and X_2 the spending an additional unit of X_3 is associated with β_3 units of Y.

If one had made three simple regressions and a single multiple regression model, the coefficients of the different X_j wouldn't necessarily be the same. This is because in the *multiple regression setting* the coefficient for one predictor represents the average increase in sales if the other predictors are held constant. This means that the effects of correlated predictors will be evened out, e.g. in a *simple regression* ice-cream sales would be associated with shark attacks, while in *multiple regression* where we add a predictor stating how many people are in the water, the effect of ice-cream sales on shark attacks will likely dissapear.

2.2 Is at least one of the predictors $X_1, X_2, ..., X_p$ useful in predicting the response?

We have to test if all the regression coefficients are zero.

$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$

Versus the alternative

$H_a : at\ least\ one\ \beta_j\ is\ non - zero$

This hypothesis test is performed by computing the *F-statistics*.

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

If there is no relationship, you would expect F to take on a value close to 1. If there is a relationship you would expect $E\{(TSS - RSS)/p\} > \sigma^2$. How large the F-statistic needs to be depends on the values of n and p. The significance (p-value) of the F-statistic is usually looked up for different n and p in a look-up table, which has values for the different F-distributions.

If we want to test the hypothesis that a subset of the coefficients are zero, we just fit another model that only uses that subset of variables.

The t-statistic and p-values for each model reports the *partial effect* of adding that variable to the model. That is, models are created with and without the variables, and their F-statistics are calculated in the following way:

$F = \frac{(RSS_o - RSS)/q}{RSS/(n-p-1)}$

Here RSS_0 is the value for the model without the variable, and q is the number of variables (1).

2.3 Do all the predictors help explain Y?

When computing the F-statistic p-value we determine if *at least one variable* is associated with the outcome, but now it is time to find which ones.

One could simply look at their p-values, but if p is large, one is likely to make false discoveries. What one instead should do is *variable selection*. One could try out all possible models and then judge their quality using one of the following: *Mallow's C_p* , *Akaike information criterion (AIC)*, *Bayesian information criterion (BIC)*, and *Adjusted R^2*.

It is, however, usually practically impossible to make all 2^p models, so instead usually does one of the following things:

- Forward Selection
    1. Start with a *null model*
    2. Fit one model with each predictor, and add the predictor with the lowest RSS
    3. Repeat 2. until some stopping rule
- Backward Selection
    1. Start with all variables
    2. Remove the variable with the largest p-value
    3. Fit a new model
    4. Repeat 2. and 3. until a stopping rule is fulfilled
- Mixed Selection
    1. Start with a *null model*
    2. Fit one model with each predictor, and add the predictor with the lowest RSS
    3. Remove a variable if its p-value has gone over the level of significance.
    4. Repeat 2. and 3. until some stopping rule

2.4 How well does the model fit the data?

2.5 Given a set of predictor values, what response value should we predict, and how accurate is out prediction?