Paul Martins

Linear Regression

Notes from ISLR book

Contents

1. Model

1.1 Principles

The goal is to find an estimate $\hat{y}$ of the variable y with a linear combination of p predictors $x_1,\dots,x_p$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{1}$$

The estimation of the intercept $\hat{\beta}_0$ and the slopes $\hat{\beta}_{1\dots p}$ is done by the least square method which minimises the Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n} \Big[y_i - (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_p x_{ip}}_{\hat{y}_i})\Big]^2 \tag{2}$$

The average of the RSS calculated over the n points is called the Mean Squared Error (MSE)

$$MSE = \frac{RSS}{n} \tag{3}$$

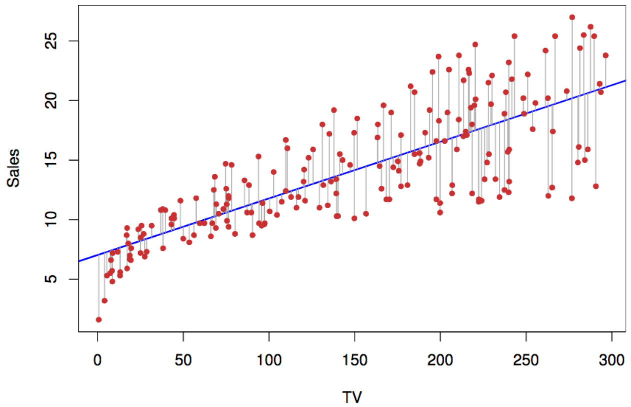

Figure 1. Linear regression fit ($p=1$), y=Sales and x_1=TV

The presence of a random noise term ε in the true relationship between x and y variables implies that the true population mean μ can only be approximated by the sample mean $\bar{y} = \hat{\mu}$. Assuming that the observations are uncorrelated, the standard error of $\hat{\mu}$ is calculated from the variance of the noise parameter $Var(\varepsilon) = \sigma^2$ which gives $SE(\hat{\mu}) = \sigma^2/n$.

Most of the time σ is not known, but we can estimate it with the Residual Standard Error (RSE)

$$RSE = \sqrt{\frac{RSS}{(n-p-1)}} \sim \sigma \tag{4}$$

This estimate of σ is also used to calculate the standard errors on the parameters $SE(\hat{\beta}_j)$[1]. Those errors are needed to get the confidence intervals $\hat{\beta}_j \pm \alpha \cdot SE(\hat{\beta}_j)$ which measure the uncertainty on coefficients[2]. Not to be mistaken with the prediction interval that quantifies the expectation on the value of a data point (ie error bands).

1.2 Response-predictor relationship

To test for a relationship between the response and the p predictors, we have 2 choices:

- t-statistic to test how far from 0 each $\hat{\beta}_j$ are, in terms of standard errors

  $$t = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} \tag{6}$$

  The p-value is the probability of finding a value equal or bigger than $|t|$. It is calculated by integrating the t-distribution with $n-2$ degrees of freedom from $|t|$ to ∞ as in Figure 2 (left). The smaller it is, the less likely we are to find a $\hat{\beta}_j$ that far from 0, therefore the higher the chance of a relationship between the response and the jth predictor.

  Note however there is always a 5 % chance of observing a p-value below 0.05.

- F-statistic to test that at least one predictor in a subset q of p predictors is linked to the response, ie at least one of the $\hat{\beta}_q$ is non-null, where $q \in [1,p]$.

  $$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \tag{7}$$

[1] In the case where $p=1$, it gives

$$SE(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right],\ SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{5}$$

[2] For the 95 % confidence interval, α is the 97.5 % quantile of the t-distribution with $n-2$ degrees of freedom (the t-distribution converges toward a gaussian as n increases).

where RSS_0 is the RSS of the model where only the $p-q$ predictors are used[3].

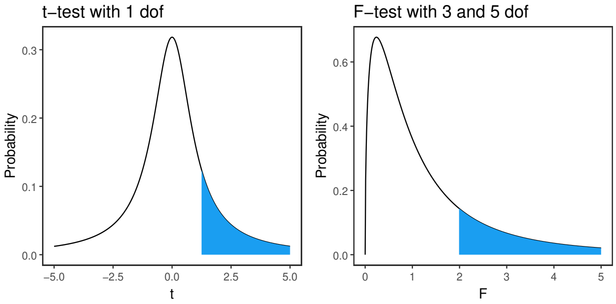

Figure 2. Example of t (left) and F (right) statistics.

1.3 Miscellaneous

Categorical predictors When a predictor is qualitative, a baseline is defined by default. In the absence of interaction terms with categorical predictors, the category effect is independent of other predictor values.

Additive assumption The effect of a change in x_i on y is independent of other predictors. Adding interaction terms will break this assumption.

Hierarchical principle Always include the main effects of the predictors that are involved in interaction terms, even if their p-values is big.

Linear assumption A change of y due to one unit change in x_i is constant regardless of the value of x_i. Adding polynomial terms will break that assumption.

2. Diagnostics

2.1 R^2 measure

A first measure of the model accuracy in the RSE. It measures the lack of fit of the model to the data but is expressed in measure of y therefore is not general to every datasets. A better metric is the R^2 test

$$R^2 = 1 - \frac{RSS}{TSS} \tag{8}$$

TSS is the total sum of squares $TSS = Var(y) = \sum_{i=1}^{n}(y_i - \bar{y})^2$. It represents the total amount of variance in the response previous to the fit.

RSS is the amount of variance that is left unexplained once the fit is performed.

R^2 represents the proportion of variance explained by the fit. The closer to 1, the better the fit !

2.2 Residual plot

Using a linear model implies that your data seems linear. A good way of checking this is by plotting the residuals vs the fitted values: it should be constantly centred around 0 like in the right top right panel of Figure 3. This residual plot also allows to check for non-constant variance of the error terms, or heteroscedasticity, by looking for a funnel shape. This is important to ensure that $\sigma^2 = Var(\varepsilon)$ is constant since the calculation of confidence intervals and errors rely upon it. Taking the log of the response or, if the response is an average over n_i values, using a weighted mean instead could help ($\sigma_i^2 = \sigma^2/n_i$), see bottom panels of Figure 3

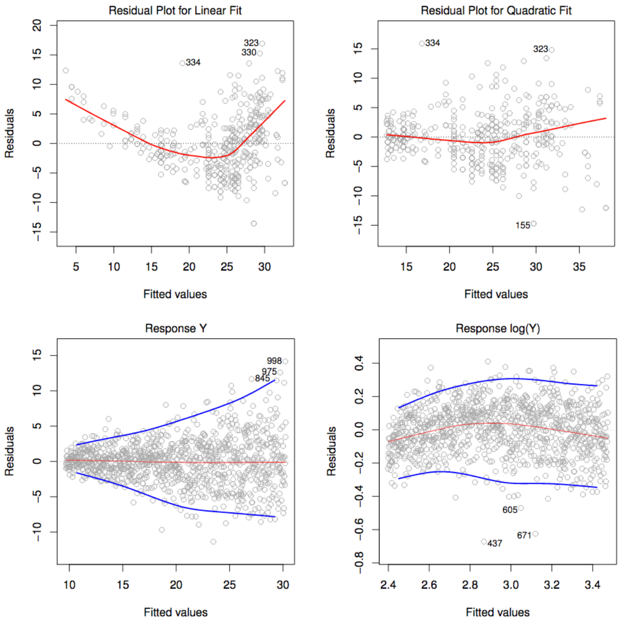

Figure 3. Top: Residuals vs fitted values for linear (top left) and quadratic (top right) regression. Bottom: illustration of heteroscedasticity.

2.3 Tracking

$SE(\hat{\beta}_j)$ are calculated assuming uncorrelated errors. If there is correlation, $SE(\hat{\beta}_j)$, and therefore the confidence intervals, will be underestimated. Imagine all data is duplicated, we have a sample of size $2n$ instead of n, same predictions but confidence interval will be narrower by a factor $\sqrt{2}$. Such correlations often occur in time series, check for tracking in the residuals as in Figure 4.

2.4 Outliers and High-leverage

- High-leverage: observation with an unusual x_i. This impact a lot the regression fit. These points are identified with the leverage statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_i)^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{9}$$

[3] See this tool for a live p-value calculation demo with t and F distributions.

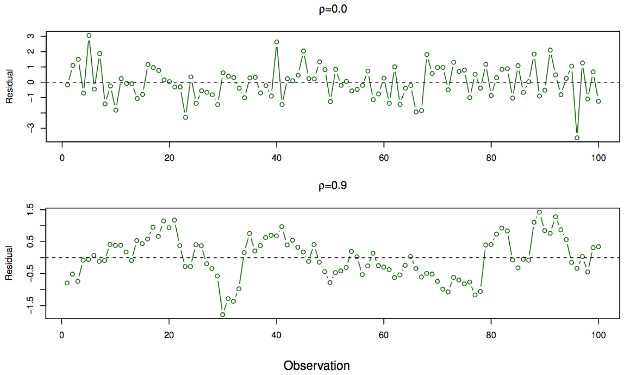

Figure 4. Top: no tracking (small correlations), Bottom: tracking (high correlations)

- Outliers: observation for which y_i is far from $\hat{y}_i$ given x_i. Outliers tend to increase RSE, hence $SE(\hat{\beta})$ and p-values. To determine these points several measures are possible such as the studentized residuals[4]:

  $$t_i = \frac{y_i - \hat{y}_{(-i)}}{\sqrt{MSE_{(-i)}(1-h_i)}} \tag{10}$$

  where the index $(-i)$ denote the value for the ith point calculated from a model where it was removed.

Both outliers and high-leverage can be spotted with the Cook's distance which measures how much all of the fitted values change when the ith observation is removed.

$$D_i = \frac{\sum_{i'=1}^{n}(\hat{y}_{i'} - \hat{y}_{i',(-i)})^2}{p \cdot MSE} = \frac{(y_i - \hat{y}_i)^2}{p \cdot MSE}\left[\frac{h_i}{(1-h_i)^2}\right] \tag{11}$$

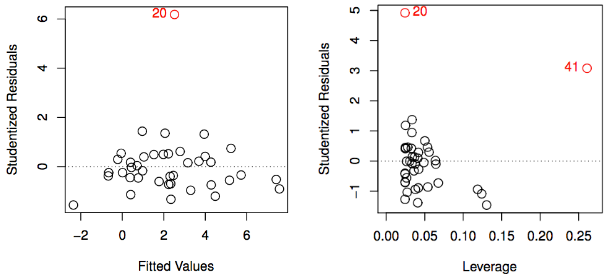

Figure 5. Studentized res.(left) and leverage (right)

2.5 Collinearity

If two predictors are collinear, it will be hard to distinguish their individual effects. This will increase the standard errors of their coefficients $SE(\hat{\beta}_j)$ and lead to a poor t-test. To detect direct collinearity, one can just look at the correlation matrix between all predictors.

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{12}$$

However, in the case of multicollinearity, it's better to look at the variance inflation factor:

$$VIF = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{13}$$

where $R^2_{X_j|X_{-j}}$ is the R^2 of a regression of X_j onto all the other predictors. It is equivalent to the variance of $\hat{\beta}_j$ calculated with a model containing all the predictors divided by the variance of $\hat{\beta}_j$ in a model with only X_j. $R^2_{X_j|X_{-j}}$ close to one means VIF high and presence of collinearity.

3. KNN Regression

If not sure about the linearity of the data, it is possible to use a non-parametric regression approach,

$$\hat{f}(x_0) = \frac{1}{K}\sum_{x_i \in \mathcal{N}_0} y_i \tag{14}$$

This could work better than linear regression if the number of predictors remain low (curse of dimensionality).

[4] NB: Studentized residuals do not consider the i^{th} point while the standardized residuals do use the full dataset.