Identification of the revenues vs. downloads characteristics of several web objects

Festo Owiny, Mat.174213
festo.owiny@studenti.unitn.it

I. Introduction

Future predictions is among the most challenging tasks organizations undertake and those that have a better handle on the subject will on average outperform others. This challenge occurs in diverse fields such as business, medicine, astrophysics, and public policy among others. Better performance can be achieved through improved analysis of data and learning relationships and structure from such data. Organizations looking for a more analytic approach to forecasting and planning revenue should consider employing leading indicators as a method to improve insight into future performance.

In this report we aim to determine relationship between two random variables by estimating the set of parameters which define a model and the corresponding noise affecting revenue measurement. In addition, we analyze the stochastic behavior and make plots of the interpolation polynomials. The strategy behind our project is to employ supervised statistical learning tools. In our case, Revenue is the response/target, Y, that we intend to predict whereas the downloads are the features/predictors, X. We write our model as; $Y = f(X) + \varepsilon$. The motivation is that with a good f we can make predictions of Y at new points $X = x$.

We hereby propose a linear regression approach for model based predictive analysis on the download data for defining most effective parameters that impacts on revenue. Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f. In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y)

II. Theory

Linear regression is a simple approach to supervised learning for predicting quantitative response Y on the basis of predictor variable X. It assumes that there is approximately a linear relationship between X and Y. linear models allow for relatively simple and interpretable inference, but may not yield as accurate predictions as some other approaches. In contrast, some of the highly non-linear approaches can potentially provide quite accurate predictions for Y , but this comes at the expense of a less interpretable model for which inference is more challenging. This function helps us answer numerous questions such as;
Is there a relationship between web downloads and revenue? How strong is the relationship between web downloads and revenue? How accurately can we predict future revenue? Is the relationship linear? Is there synergy among different predictors if any? We assume a model,

$$Y = \alpha X + \beta + \varepsilon$$

Where α and β are two unknown parameters that represent the intercept and slope respectively, and ε is the irreducible error. Given some estimates $\hat{\alpha}$ and $\hat{\beta}$ for the model coefficients, we predict future responses using $\hat{y} = \hat{\alpha}x + \hat{\beta}$.
When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are; non-linearity of the response-predictor relationships, correlation of error terms, non-constant variance of error terms, outliers, high-leverage points, and collinearity.

A. Least squares

Least squares is a standard approach in regression analysis to approximate linear models. The goal is to obtain estimates $\hat{\alpha}$ and $\hat{\beta}$ such that the linear model fits the available data. The most common approach of fitting involves minimizing the least squares criterion. From $\hat{y_i} = \hat{\alpha}x_i + \hat{\beta}$, the prediction for Y based on the i^{th} value of X; implies $e_i = y_i - \hat{y_i}$ represents the i^{th} residual and we define the residual sum of squares (RSS) as $RSS = e_1^2 + e_2^2 + ... + e_n^2$; n = number of observations. The least squares approach chooses $\hat{\alpha}$ and $\hat{\beta}$ to minimize RSS s.t.;

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

B. Accuracy

The standard error (SE) of an estimator reflects how it varies under repeated sampling. We have;

$$SE(\hat{\alpha})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ SE(\hat{\beta})^2 = \sigma^2\left[\frac{1}{n} + \frac{\hat{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right],$$

Where $\sigma^2 = Var(\epsilon]$; the estimate of σ is known as the Residual Standard Error (RSE); $\quad RSE = \sqrt{\frac{1}{n-2}(RSS)}$

The SE can be used to compute confidence intervals (C.I) of parameters; $\hat{\alpha} \pm 2.SE(\hat{\alpha})$; and to perform hypothesis tests on the coefficients; Null hypothesis, $(H_0 : \alpha = 0)$ - there is no relationship between X and Y *versus* alternative hypothesis, $(H_A : \alpha \neq 0)$ - there is some relationship between X and Y. If $\alpha = 0$ then the model reduces to $Y = \beta + \epsilon$, and X is not associated with Y. Once H_0 is rejected in favor of H_A, we quantify the extent to which the model fits the data using RSE and R^2 statistic. The RSE provides an absolute measure of lack of fit of the model to the data. The $R^2 = \frac{TSS - RSS}{TSS}$; Its large value indicates that a large proportion of the variability in the response has been explained by the regression.

C. Distributions

The *Student's t-distribution* has a bell shape for values n>30; quite similar to the normal distribution. For i.i.d. population with sample mean $\hat{x}$ and sample variance s^2;

$$t = \frac{\hat{x}-\mu}{\frac{s}{\sqrt{v+1}}}$$

with v degrees of freedom. As $v \to \infty$, it converges to std.normal $t \sim N(0,1)$. Test H_0 with $t = \frac{\hat{\alpha}-0}{SE(\hat{\alpha})}$, $n-2$ degrees of freedom (*df*), assuming $\alpha = 0$. p-value is the probability of observing any value equal to $|t|$ or larger.

If $X1, X2, \ldots, Xm$ are m independent R.Vs of $N(0,1)$, then $V = X_1^2 + X_2^2 + ... + X_m^2 \sim \chi^2$ (*Chi-squared distribution*) with *df*(m), mean(m) and variance($2m$);

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

As its *df* $(n-1)$ increases, $\chi^2 \sim N(\mu, \sigma^2)$. It is also a component of definition of the t-distribution and used in t-tests, analysis of variance, and regression analysis. For a confidence level $1-\phi$ and variance σ^2, the $C.I.$ for the population standard deviation with $(n-1)$ *df* is;

$$\sqrt{(n-1)s^2/\chi^2_{\phi/2}} \leq \sigma \leq \sqrt{(n-1)s^2/\chi^2_{1-\phi/2}}$$

The *Uniform distribution* is a symmetric probability distribution where finite number of values are equally likely to be observed. It is also a discrete uniform random variable in the interval $[a, b]$. In our task $a = 0$ and $b = D_i^{max}$.

III. Implementation

In this section we explain the work-flow taken up in the implementation as well as the architecture adopted. We employed R, a statistical computing package to implement data analysis because of its versatility, interactivity, and popularity. Initially, with the help of R functions; *summary()*, *ggplot()* and others, we analyzed our data-set in order to verify the structure, format and detect any errors or missing values. We then conceptually split the data-set into 10 chunks $w.r.t$ to the web objects in order to estimate D_i^{max}; values obtained by maximizing the differences $(x_i(t+1) - x_i(t)) \; \forall \; t = 1, ..., 2400$ and $i = 1, ..., 10$. To estimate the parameter of the noise affecting the measure of revenues, σ, we averaged RSE values of the 10 web objects. Since the original population of data is normally distributed, we use chi-square distribution to construct confidence intervals for the variance and standard deviation with $\chi^2 - distribution$ at confidence level 99.7% and df 2399 calculated the critical values and confidence interval for standard deviation σ respectively.

The model implemented excludes the intercept implying $\beta = 0$. Therefore, the main mathematical framework (model) adopted is;

$$Y = \alpha X$$

This model has slope α and passes through the origin. Furthermore, we used *lmList()* function to fit regression models $w.r.t$ the ten web objects then investigated the model properties such as slope, RSE and $p-$values. In addition, *abline()*, *plot()* and *ggplot()* functions are employed to plot the interpolation polynomials. Given parameter estimates $\hat{\alpha}_i$ we make predictions $\hat{y}_i$ corresponding to the models on the basis of $X = x$; $\hat{y}_i = \hat{\alpha}_i x$
These predictions are then deducted from the actual response values to obtain the residual errors after which several plots made.

IV. Evaluation

Here we show assessment results from data, predicted model and the associated noise. The both box plots are generally symmetrical about the medium and comparatively tall signifying increasing downloads and revenue respectively with anomalies in 1 and 5 of (b) and (c)that tend to drag below others. While web object 4 generated the highest revenue, 1 and 5 generated least. The point plot (a) also shows close similarity in the linearity of the data across all the web objects. Furthermore, these plots show no outliers as below.

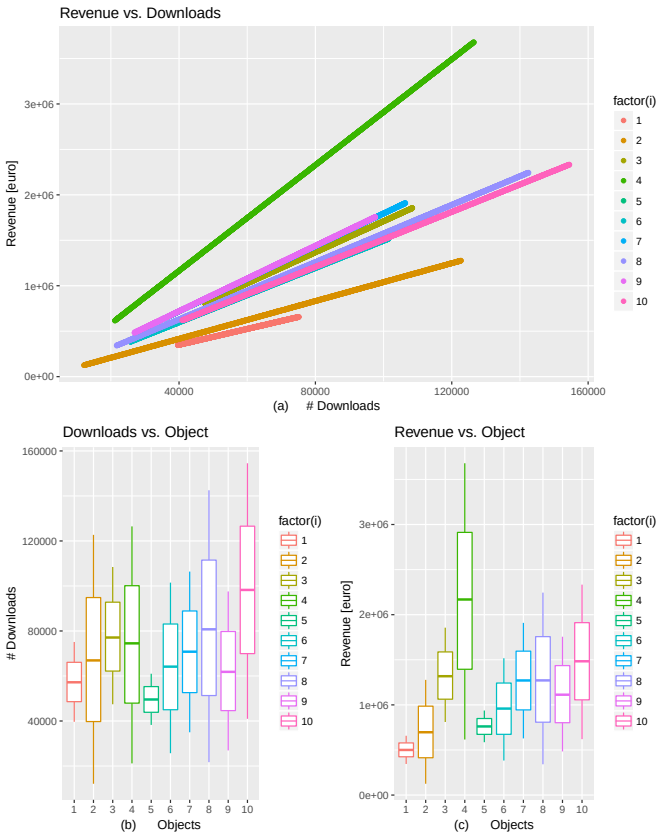

Figure 1. Point plot (a) & box plots (b) and (c) to assess the data

Table 1 shows that the greatest download difference () are in web objects 4, 2, 10 and 8 respectively; while 3,6,7 and 9 lie in close interval. Likewise slopes α_i lie within a small interval with largest difference between 1 and 4.
$RSE = 1271.119$ on 23990 degrees of freedom.
Hence Standard deviation, $\sigma = 1271$
The 99.7% Confidence interval for σ is [1218 , 1327]

i	1	2	3	4	5	6	7	8	9	10
RSE_i	1271	1289	1278	1298	1261	1258	1255	1269	1259	1272
α_i	8.75	10.40	17.10	29.10	15.35	14.95	17.95	15.75	18.00	15.10
D_i^{max}	30	92	50	87	19	63	59	99	59	94

Table.1. Showing estimated parameter sets for the 10 web objects.

Figure 2 shows close similarity in the linearity among the models. Largest difference is realized between model 4 and 1. There is close similarity in the other models. This behavioral pattern is also observed in α_i values of Table 1.

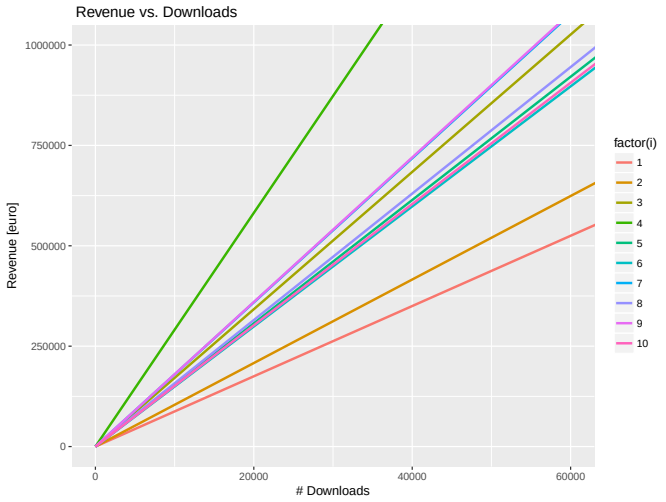

Figure 2. Linear Interpolation polynomials corresponding to estimated model parameters

Figure 3 (a) is symmetrical with no discernible pattern hence uncorrelated errors and evidence for linearity of data. Correlations among the error terms frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time; as shown in (b) also is symmetrical and shows no discernible pattern thus uncorrelated errors.

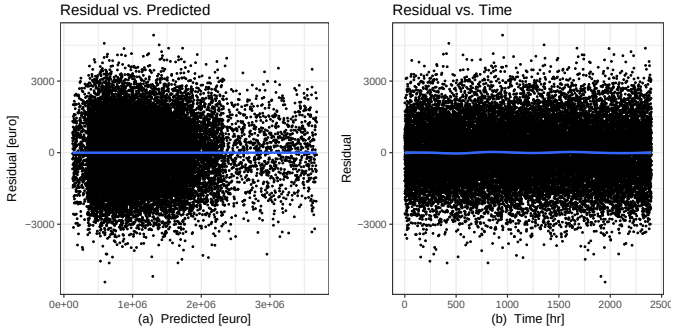

Figure 3. Residual plots. In each, the blue line is a smooth fit to the residuals with no pattern in the residuals; correlation = -0.00064

The `S.Es` and `C.Is` associated with our models rely upon the assumption that error terms have a constant variance. The 99.7% Confidence Interval for `S.D`, σ is narrow [1218 , 1327] and residual plots are symmetrical with constant variance and very minimal outliers thus full support of the model.

V. Conclusion

The evidence of the assumption that noise is stochastically same for all objects improves the reliability of our model. This demonstrates a strong linear relationship between downloads and revenue and therefore future predictions can be accurately made. The simplified linear model is a good choice for our data but we could further investigate download diversity over a large scope of web objects such that multiple features are retrieved from it to be implemented in a multiple linear regression model for consistent accuracy over a wider scope. Finally, the assumption of uncorrelated errors is extremely important for the model yet it could be violated as correlation may occur outside of time series data, therefore a good experimental design is crucial in order to mitigate the risk of such correlations.

References

[1] K.S. Trivedi. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. , 2001.

[2] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. , 2013.

[3] Hadley Wickham. *ggplot2, elegant graphics for data analysis*. , 2016.