MIXED - METHODS MMM

The Utilization of Ridge Regression in Marketing Mix Modeling

Prof. Dr. Alexander Preuss[1] Antonio Prada[2]

Abstract

Since the introduction of Marketing Mix Modeling (MMM) in the advertising industry, it has undergone a significant evolution, transitioning into a more intricate measurement solution over time. Initially grounded in established econometric methodologies originating from academia, MMM has incorporated specific advertising research techniques, including variable transformation and the modeling of saturation and carryover effects, which are now integral components. Presently, the landscape of MMM confronts ongoing challenges such as the emergence of diverse digital platforms and the management of extensive datasets, coupled with current issues related to signal resiliency and reduced tracking possibilities. As a result, the domain of MMM is increasingly influenced by data science, featuring the integration of automation and machine learning (ML) techniques. Many practitioners have shifted from creating basic Ordinary Least Squares (OLS) models to more advanced and multifaceted models. Within the realm of frequentist modeling, the utilization of contemporary statistical and ML-related methods, involving techniques such as regularization methods (L1 - LASSO and L2 - Ridge), has gained traction to address issues with multicollinearity due to media fragmentation. However, using these methods alone may not consistently produce robust model results concerning crucial variables represented in a model. In this paper, we propose a mixed-methods approach to MMM, combining traditional OLS methods with contemporary statistical and ML-related elements, such as ridge regression, through a two-stage modeling process.

[1] Professor of Quantitative Methodology, Fresenius University of Applied Sciences, Hamburg, Germany and Managing Partner Marketing Science (bynd GmbH, the strategy and technology consultancy of Omnicom Media Group)

[2] Group Manager Marketing Science (Germany), Snap Inc.

1. Introduction

With media fragmentation on the rise and challenges in ad tracking, exemplified by Apple's introduction of ATT/App Tracking Transparency in 2021 and Google's plan to end third-party cookies in 2025, advertisers are increasingly turning to signal-resilient solutions like dynamic and automated Marketing Mix Modeling (MMM) methods. In addition to partnering with agencies, vendors, and SaaS providers, an increasing number of advertisers opt to develop their own automated in-house MMM solutions. Furthermore, open-source MMM codes are gaining in popularity: Notable examples include Meta's Robyn (Zhou et al., 2021) and the recent Google Meridian launch (March 2024) based on its previous MMM open-source code (Jin et al., 2017).

Automated MMM methods driven by machine-learning (ML) techniques allow modeling practitioners to handle an increasing number of variables in models, mostly required due to emerging digital platforms and the demand for more granular modeling results. However, in order to reduce issues of overfitting and multicollinearity, many modern MMM systems prioritize contemporary regression methods enhanced by machine-learning approaches and favor high-level modeling results with respect to goodness of fit (R squared) over more conventional methods that typically ensure statistically robust results across modeled sets of variables. In some cases, complex built-in features and specific methods can mitigate these effects but require extensive resources. In this paper, we propose a less complex but powerful approach to effectively model and break out online channels by focusing on established statistical methods.

Bynd Consulting (OMG Germany) has tackled this issue and developed an innovative modeling approach which is already the standard approach in its daily business. This method produces conventional industry-standard MMM outputs (OLS regression) and then in a second-stage builds and uses ML/Ridge regression techniques to quantify online channels including social media platforms within the constraints of the main reference model (two-stage modeling process). This method is statistically more robust than other MMM approaches that use mainly ML-only methods in a direct modeling approach.

2. Problem Statement and General Approach

As media continues to fragment and new platforms emerge, it becomes necessary to quantify smaller impact factors in marketing mix modelings. This is where classical modeling approaches reach their natural limits. The number of observation points determines the number of detectable independent variables in a model. In the case of multiple linear regression, the number of variables cannot exceed the total number of observation points. If other statistical indicators such as the significance of the regression coefficients used in the model are taken into account, then the possible total number of variables in a model is much more limited – as a rule of thumb, one variable for every ten observation points can be used in a model. Another essential prerequisite is that the effect of the variables in the "real world" must be sufficiently large to be detectable in the model, which always represents a simplification of reality. This makes it even more difficult to include smaller measures in the model in a meaningful way and the addition of several smaller platforms could lead to severe multicollinearity issues.

Classical Marketing Mix Modeling (MMM) is typically applied to weekly data (in most cases, weekly data for the past three years is usually available) from which up to 15 variables can typically be detected according to this rule of thumb. The main problem now is how to detect and quantify smaller impact measures in a model while taking important statistical criteria into account, such as the significance of the variables used in a model.

Technically, this whitepaper will discuss a novel modeling approach of combining modern elements such as methods popular in machine-learning processes and conventional statistical methods. In addition, we will utilize the proposed two-stage modeling method and conduct an accompanying database analysis based on 17 sales model cases. These cases range across 4 industry sectors (beverages, pharmaceuticals, FMCG, retail), with annual gross total media spend ranging from 2.8m to 109m in the German market. The main focus will be on quantifying online variables represented by buckets (social media vs other digital media buckets such as video, social, and search/SEA) while other non-digital media variables will be modeled in the first stage and function as control variables in the second stage of the modeling process. The results obtained through the two-stage modeling process will be compared to models using primarily OLS methods and other statistical methods associated with ML-related approaches (ridge regression). By comparing modeling results and their accompanying statistical criteria, we should be able to select the superior model in the process. More clearly, since the OLS model (stage 1) will function as the reference model, we can obtain the quantified online variables by applying ridge regression techniques (stage 2) by creating an auxiliary model in the process, and then link the modeled individual digital channels to the pre-modeled total digital bucket in the original reference (OLS) model.

In this study, we will test this proposed modeling method by applying actual sales data of an unspecified client within the FMCG sector in Germany. Several media channels were used to advertise the product, including TV along with individual digital channels such as online display, addressable TV (ATV), online video, and social media. In this example, the data is anonymized and partially subjected to transformations.

3. Proposed Method: Two-stage Mixed-Method MMM

Since the popularization of MMM in the advertising industry, it has evolved significantly over time. MMM mostly relies on established econometric methods while incorporating specific advertising research techniques. The most popular approaches involve Bayesian and frequentist methods, commonly multiple regressions such as OLS and more recently regularization techniques (L1/L2). However, due to growing challenges involving the fragmentation of different advertising channels (especially digital platforms ranging from social media channels to other online services), traditional OLS methods are now no longer sufficient in breaking out all channels in the process. On the other hand, more advanced techniques involving regularization techniques are not fully compatible with the established consensus on how to build and utilize marketing mix models with respect to ensuring statistical robust results across impact factors. Therefore, it would be of interest to explore the possibility of combining the best available techniques: Since traditional OLS methods provide relevant information on the statistical significance of individual variables, it would be essential to continue to rely on this method. However, with respect to smaller advertising channels (mostly digital advertising channels), we would need to introduce another technique allowing us to break out these channels more effectively. We could specifically use regularization techniques such as ridge regression to achieve this goal. Now by combining these two different methods, we could link traditional methods such as OLS with more flexible and contemporary methods related to regularization techniques. This combination of methods leads us to the proposal of using “mixed-methods" in two subsequent modeling stages when conducting modern marketing mix modelings (Fig. 1).[3]

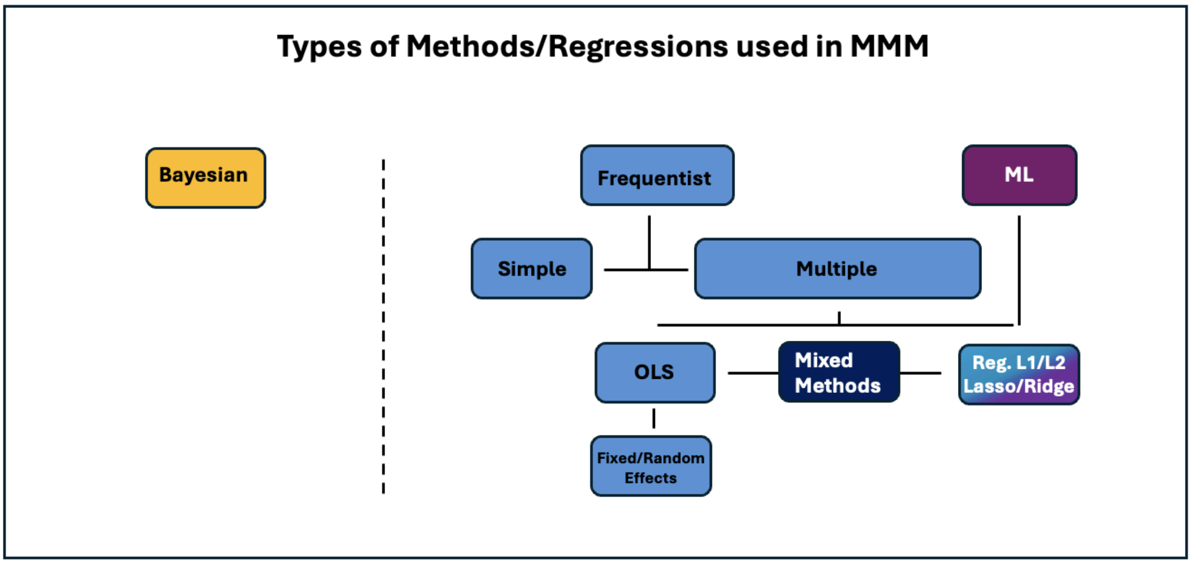


Fig. 1: Types of Methods/Regressions used in MMM

The purpose of this section is to show how two different statistical approaches can be combined in the realm of marketing mix modeling. "Classical" multiple linear regression is a relatively established method and used to explain an independent

[3] The various methods shown in Fig. 1 are not exhaustive and could be extended by adding more techniques such as random forest, among other methods. In addition, MMM’s can be more complex with more added depth including hierarchical modeling approaches.

variable by applying a linear combination of independent variables. The major advantage of this process is its robustness and complete transparency. Through statistical tests, it is possible to evaluate the "quality" of the model, including the "confidence level" (statistical significance) of the used variables. This is represented through their respective regression coefficients, in order to avoid taking into account unrelated “impact factors" that would attract variance and thus distort the overall model. This is simultaneously the greatest strength and also the greatest weakness of this approach. The high level of statistical confidence is offset by the limitation of numbers of explanatory variables used in the model. This means the model can only consist of a limited number of variables. This number depends on the number of data points used in the model – as a rule of thumb (empirical value), a number of variables of data points divided by 10 can be included and quantified in a model, i.e. roughly 15 variables in case of weekly data which are available for a period of three years (= 156 data points). Variables that cannot be detected in a model are usually measured with respect to smaller impact due to lower levels of investment. However, since the detection of smaller digital media channels is of particular interest, a comprehensive method is required that enables a high statistical confidence of the modeled results while it extends the limitation of the number of variables used in a model when compared to "classical" regression. One such approach would be the combination of "classical" regression with "ridge regression" methods in a two-stage modeling approach.

This proposed Two-Stage Mixed-Methods MMM approach is illustrated in the figure below (Fig.2):

- Stage 1: In the first modeling stage, a traditional OLS model is created. Individual digital channels are combined in a “total digital bucket”. This model functions as a reference model.
- Stage 2a: In an independent second modeling-stage, a conventional ridge-regression model is created. This auxiliary model contains the exact same data as the original OLS model but the individual digital channels are broken out in this model.
- Stage 2b: Finally, we can combine both models by using the modeled results of the non-digital variables from the OLS model and the modeled individual digital variables from the ridge-regression model. However, we cannot use these modeled digital variables directly and will have to weigh each individual variable against the pre-modeled digital total variable (reference model) accordingly.

This approach ensures that modeled individual variables are required to meet statistical significance (here: .05 level), either directly or indirectly via model output weighting. In the alternate ridge-only approach, this is not the case and the goodness of fit (R squared) is the main criterion to determine model superiority. In addition, issues with selecting the right lambda leads to the creation of biased models (in exchange for variance) in the process. In contrast to other regularization methods (e.g. LASSO), ridge regression methods do not perform variable selection for all predictors in the model [see detailed discussion in sections 3(b) and 3(c).

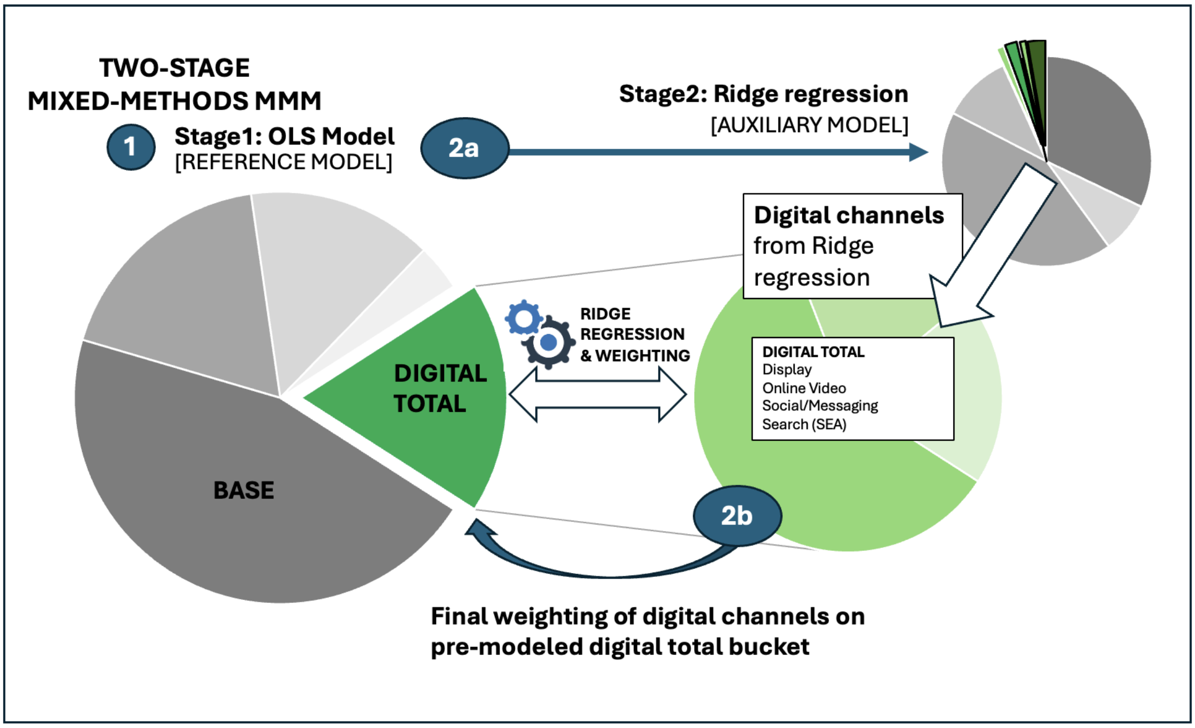


Fig. 2: Concept of Mixed-Methods MMM

a. "Classical" Modeling

"Classical modeling" uses multiple regression to estimate the influence of the independent variable on the dependent variable, e.g. total sales revenue (Y). This process is relatively established, it was first mentioned at the beginning of the 20th century (Pearson, 1908). However, nowadays, updated methods are being used when applying this approach in order to take the carryover effect of media into account or to map how the effect changes depending on the “ad pressure” level.

The general objective of the model creation is to explain the target variable (e.g. sales over time) through a combination of independent variables. Multiple regression is a statistical technique that determines "optimal weight factors" for a combination of independent variables that best explain the dependent variable. For example, sales = x1*price + x2*temperature + x3*ad spend (the xi are the weight factors or model coefficients). In practice, the main goal is to find an "ideal combination" of independent variables that provide a good fit to the dependent variable in a statistical sense, but at the same time are plausible in terms of content. The model quality (R^2, goodness of fit) indicates what proportion of the dispersion of the target variables is explained by the combination of the model variables (Preuss, 2019).

In order to illustrate the results of this procedure, a model for an FMCG is presented as an example. The analysis considered the sales of an FMCG advertiser on a weekly basis[4]. The digital variables Display, ATV, Online Video and Social were initially combined into a common variable "Digital total". For this purpose, the weekly ad impressions were summed up into a total variable.

Furthermore, additional control variables were used to represent the different lockdown periods during the 'Covid-19' pandemic phase. For this purpose, fundamentally different variables might be suitable, such as mobility data or data on the number of infections (number of confirmed cases). However, in many models, such as this one, a simple structural dummy variable D (D = 1, during the period, D = 0, outside the period) has proven to be effective when modeling post-pandemic. This variable expresses an additional positive additive measure in the model function during the lockdown period and thus helps to capture the effect of the other variables more realistically and also more reliably in a statistical sense.

In order to take the time-lag on advertising effect into account, a procedure based on the "ad stock" method (Broadbent, 1979) was used in this case. As a result, in this specific case, a carry-over rate of 96% was determined for TV, a corresponding rate of 50% for online, i.e. 4% of the effect of TV is attributable to the first week. The remaining 96% with decreasing intensity is attributed to the following weeks, while for online, 50% of the effect already unfolds in the current week (i.e. the carryover-effect is significantly lower than for TV).

As a result, the following model was created for total sales units (Y):

MODEL OUTPUT – OLS REGRESSION

Coefficients	Estimate	Std. Error	t value	Pr (> | t|)
Baseline	38048.3	5595.1	6.8	***
Seasonality	4781	1758.5	2.719	**
Price (conv. unit)	-5186.8	640.2	-8.102	***
Distribution, Flyers (weighted)	1550.5	125	12.401	***
Distribution, Flyers/Display (weighted)	2266.1	197	11.504	***
Shopping days	17891.1	1790.2	9.994	***
Covid19 (Lockdown 1)	17255.5	4278.4	4.033	***
Covid19 (Lockdown 2)	11110.8	2869.2	3.872	***
TV (GRPs) log adstock 96	2397.4	1193.4	2.009	*
Digital total (AIs) log adstock 50	685.5	194.8	3.52	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R^2 = 0.9181
Durbin-Watson = 1.8305

Table 1: Model statistics, OLS regression

Key indicators for assessing the model are the multiple R^2 and the probability of error of the regression coefficients. Details on this can be found, for example, in Bortz and Schuster (2016).

In the original model (Fig. 3), 91.8% of the dispersion of the target variables can be explained by the combination of the detected variables (R^2 - goodness of fit) – this is a relatively good value when conducting classical MMMs. At the same time, it is shown that all variables used in this model are statistically significant ($p < 0.05$).

[4] For this anonymized example, a linear transformation was applied to the variables.

The Durbin-Watson statistic (DW = 1.83) is near the required critical value of DW = 2.0 which shows that the residuals are [nearly] free of auto-correlation. This leads to the conclusion that no essential variables are missing in this model. For more information on the Durbin-Watson statistics, see Durbin and Watson (1971).

The graphical representation of the model also shows the high degree of model fit (Fig. 3): The actual total sales revenue (dotted black line) is captured very well by the model (red line). Accordingly, the residuals, i.e. the difference between the values of the real sales and the model values (blue line), are correspondingly small. In addition, there is no structure to be seen in the residuals, which is already indicated by the Durbin-Watson statistic.

Based on the model, key results such as sales decomposition or return on investment (ROI/ROAS) can be calculated, but there is no uniform standard for this (cf. Preuss, 2019). The sales decomposition shows the extent to which the variables detected in the model explain the target variable. In addition to the modeled impact drivers, the sales decomposition usually also includes a so-called "baseline", which describes the proportion of the target value that is not attributable to the impact factors. These can be, for example, long-term effects of communication, positive experiences or – in the case of sales – repeat purchases. In this respect, the baseline can also be interpreted as "brand strength".

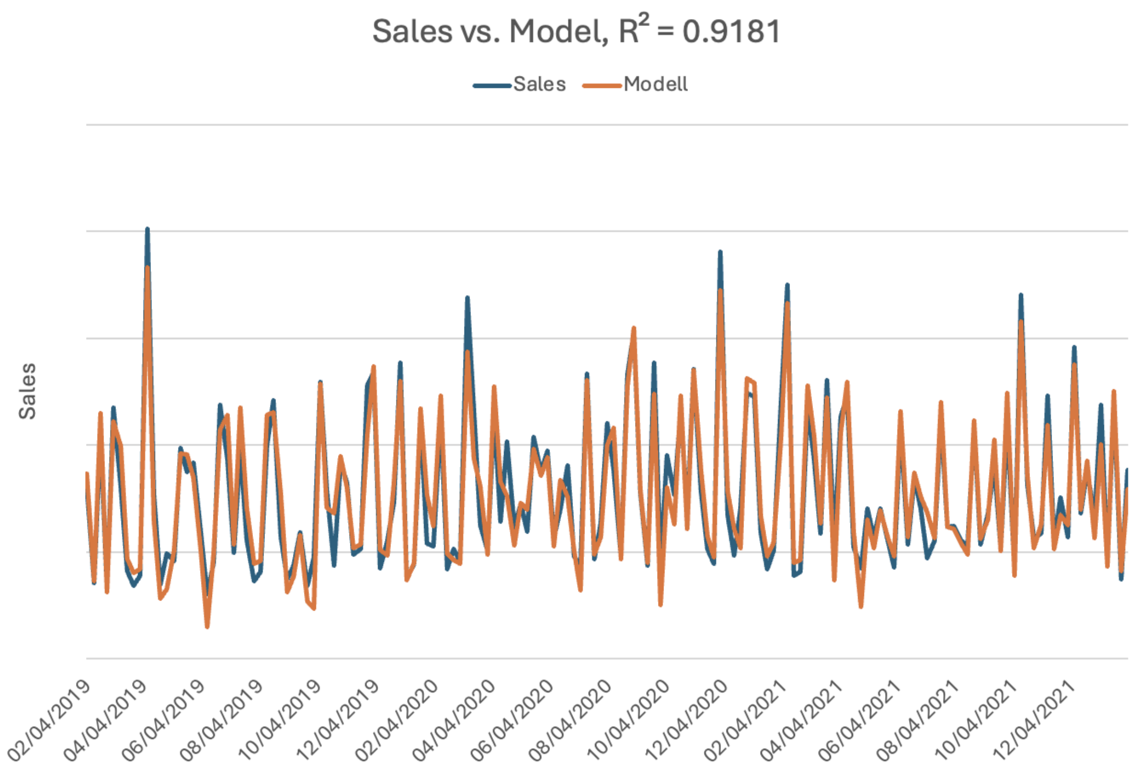
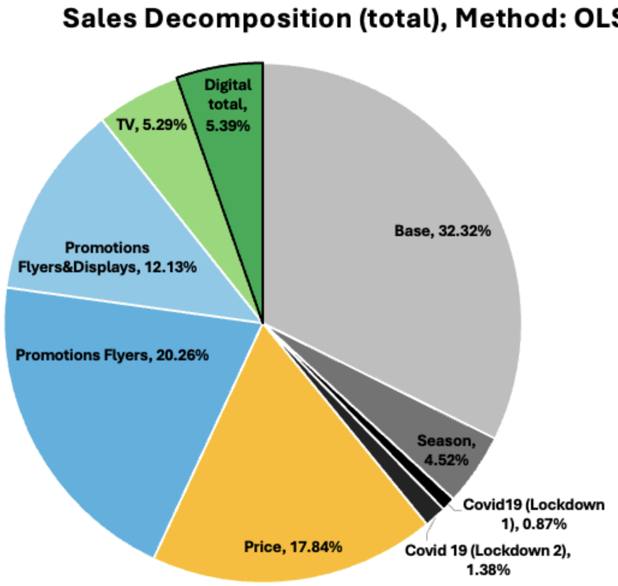


Fig. 3: Target variable (actual sales, Y) vs. model fit

In this example, the decomposed model shows that price explains 17.84% of unit sales, 5.29% is based on TV commercials, and 5.39% based on total digital ads.



Key Driver	Share	Method
Base	32%	OLS
Season	5%	OLS
Price	18%	OLS
Promos	32%	OLS
TV	5%	OLS
Digital total	5%	OLS

Fig. 4: Sales decomposition based on OLS model output results

The main issue is to determine how the individual digital channels can be further quantified in the model.

For now, let's inspect another model in which the total digital variable has been replaced by its respective sub-components (individual digital channels):

MODEL OUTPUT – OLS REGRESSION

Coefficients	Estimate	Std. Error	t value	Pr (> | t|)
Baseline	36539.85	5634.88	6.485	***
Seasonality	3328.25	1982.08	1.679	
Price (conv. unit)	-4930.16	670.89	-7.349	***
Distribution, Flyers (weighted)	1553.48	126.81	12.251	***
Distribution, Flyers/Display (weighted)	2301.77	198.35	11.605	***
Shopping days	17185.36	1850.25	9.288	***
Covid19 (Lockdown 1)	17068.09	4291.15	3.978	***
Covid19 (Lockdown 2)	10529.63	2893.79	3.639	***
TV (GRPs) log adstock 96	2900.48	1323.57	2.191	*
Display (Imps) log adstock50	811.68	255.89	3.172	**
ATV (Spend) log adstock50	-200.83	418.36	-0.48	
Online Video (Imps) log adstock50	20.96	260.53	0.08	
Social (Imps) log adstock50	680.44	211.32	3.22	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R^2 = 0.9199
Durbin-Watson = 1.8847

Table 2: Model statistics, classical regression using all digital measures

The modeled output results (method: OLS) in Table2 show non-statistically significant variables (ATV, OL video). Therefore, it is not possible to take the effect of all digital channels into account at the same time (sometimes this is due to high probability of error, wrong pos./neg. signs). In fact, the lack of statistically significant results of the modeled individual variables is mainly due to multicollinearity issues.[5]

b. Approach: Ridge Regression

In general terms, ridge regression is a recent modification of classical linear regression.

Statistical background:

In linear regression, optimal weight factors bi are determined for a linear combination of independent variables that are supposed to best predict a dependent variable in terms of the least squares. The weight factors can be determined by a comparatively simple equation. In matrix notation, this equation [1] is formulated as

$$[1]\ \underline{b} = (\underline{A}'\underline{A})^{-1}\underline{A}'\underline{y}$$

which is where $\underline{b}$ is the vector that contains the regression coefficients, the "optimal weight factors" as mentioned above. Matrix $\underline{A}$ contains the vectors formed by columns from the values of the independent variables. Finally, $\underline{y}$ is the vector that contains the values of the dependent variable. For further details, please refer to a suitable textbook, e.g. Bortz and Schuster (2016).

In order for the equation to be satisfied (or more precisely, for the invertibility of matrix $\underline{A}'A$ to be possible), matrix A must contain at most as many variables as data points. In practical applications, however, the number is often substantially lower, particularly if all regression coefficients are to be statistically significant.

Ridge regression effectively removes this constraint by introducing a regulating factor λ, thus changing the equation [2] to

$$[2]\ \underline{b} = (\underline{A}'\underline{A}+\lambda\underline{I})^{-1}\underline{A}'\underline{y}$$

More details can be found in Berk (2020). This factor – if the size is appropriate – leads to the model being able to absorb more variables and has the "positive" side effect that the signs of the coefficients for the media variables of interest often become positive as λ increases. Even if the model (significantly) deteriorates when measured as a ratio such as R^2 with increasing λ compared to a model with λ = 0, it is still a "legitimate" regression in which the coefficients are

[5] Multicollinearity refers to the case where the explanatory variables in the model are too highly correlated with each other. One effect of this is that not all variables are significant at the same time. This becomes clear when you take a closer look at the structure of the test variable for the significance test. The test variable is described, for example, by Bortz and Schuster (2016). Furthermore, tests for multicollinearity such as VIF (Variance Inflation Factor) can be conducted to determine the level of multicollinearity associated with the individual variables in a model. However, these tests can only detect the presence of multicollinearity in a model but do not fix the issue directly. In addition, VIF tests should be used cautiously in practice (O'brien, 2007). More detailed discussions would go beyond the scope of this paper.

estimated on the basis of the data structures (λ is a number > 0 within ridge regression). The practical task is to find a suitable value for λ.

The strength of the coefficients also decreases as λ increases, with the ratio of coefficients remaining relatively stable above a certain value for λ. Therefore, it is a "legitimate" output result created by the regression. Generally, the variables used in ridge regressions are usually not statistically significant, although in the following examples the significance test is applied, which can no longer be meaningfully interpreted due to the properties of the ridge regression. The probabilities of error can only be used as a benchmark.

c. The Use of Ridge Regression

a) Single-stage Model

In the first step, the "Online total" variable is replaced by the individual variables "Display", "ATV", "Online Video", and "Social" in the model. The carryover rate of the total variable – in this case (adstock 50%) – will be adopted for the individual variables in order to avoid the creation of further distortions. Depending on the distribution of ad impressions, a different carry-over value for each variable can affect the regression coefficient of the variables.

In this example, λ is set to 10,000:

MODEL OUTPUT – RIDGE REGRESSION, Lambda = 10000

Coefficients	Estimate	Std. Error	t value	Pr (> | t|)
Baseline	"23.2582"	"207.0695"	"0.1123"	"0.9107"
Seasonality	"14.3582"	"206.4018"	"0.0696"	"0.9446"
Price (conv. unit)	"1634.5956"	"140.5653"	"11.6287"	"0"
Distribution, Flyers (weighted)	"2294.7868"	"124.302"	"18.4614"	"0"
Distribution, Flyers/Display (weighted)	"1435.0837"	"172.8733"	"8.3014"	"0"
Shopping days	"796.2482"	"198.4142"	"4.0131"	"0.0001"
Covid19 (Lockdown 1)	"16.7317"	"207.0311"	"0.0808"	"0.9357"
Covid19 (Lockdown 2)	"39.3334"	"206.9285"	"0.1901"	"0.8495"
TV (GRPs) log adstock 96	"357.6426"	"204.3319"	"1.7503"	"0.0822"
Display (Imps) log adstock50	"384.4832"	"174.8768"	"2.1986"	"0.0295"
ATV (Spend) log adstock50	"130.5388"	"195.1072"	"0.6691"	"0.5045"
Online Video (Imps) log adstock50	"188.621"	"174.8931"	"1.0785"	"0.2826"
Social (Imps) log adstock50	"502.6072"	"165.9804"	"3.0281"	"0.0029"
---	""	""	""	""
R^2 = "0.7685"				
Durbin-Watson = "1.5447"				
Lambda = "10000"				

Table 3: Model statistics, ridge regression, with λ = 10,000

The model R sq. now drops to 0.77, the digital variables are not statistically significant - when the t-test intended for the "classical" multiple linear regression is carried out, but all of these variables are positive in the model (Table 3). The calculated ratios of the coefficients of digital channels (index ATV = 100) are as follows:

Display: 319, ATV: 100, OL Video: 156, Social: 417

For comparison, results of regression with larger factors λ = 100,000 and λ = 1,000,000:

MODEL OUTPUT – RIDGE REGRESSION, Lambda = 100000				
Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr (> | t|)**
Baseline	"3.878"	"165.5286"	"0.0234"	"0.9813"
Seasonality	"1.1716"	"165.4682"	"0.0071"	"0.9944"
Price (conv. unit)	"892.6041"	"148.7801"	"5.9995"	"0"
Distribution, Flyers (weighted)	"1141.1405"	"143.6291"	"7.9451"	"0"
Distribution, Flyers/Display (weighted)	"532.6466"	"159.7884"	"3.3335"	"0.0011"
Shopping days	"393.7989"	"162.8736"	"2.4178"	"0.0169"
Covid19 (Lockdown 1)	"4.1599"	"165.5254"	"0.0251"	"0.98"
Covid19 (Lockdown 2)	"10.8041"	"165.5162"	"0.0653"	"0.948"
TV (GRPs) log adstock 96	"160.6901"	"164.9964"	"0.9739"	"0.3317"
Display (Imps) log adstock50	"221.1747"	"161.5634"	"1.369"	"0.1731"
ATV (Spend) log adstock50	"75.9566"	"164.1447"	"0.4627"	"0.6442"
Online Video (Imps) log adstock50	"189.4926"	"161.4571"	"1.1736"	"0.2425"
Social (Imps) log adstock50	"243.8306"	"159.9823"	"1.5241"	"0.1297"
---	" "	" "	" "	" "
R-Quadrat = "0.7362"	"0.7362"			
Durbin-Watson = "0.5832"	"0.5832"			
Lambda = "100000"	"100000"			

Table 4: Model statistics, ridge regression, with λ = 100,000

The R sq. drops to 0.74, the digital measures are - in conventional determination - not significant, but all with a positive sign. The calculated ratios of the coefficients of digital measures:

Display 291, ATV: 100, OL Video: 249, Social: 321.

If the value for λ is now 10 times larger, the following model statistics are obtained:

MODEL OUTPUT – RIDGE REGRESSION, Lambda = 1000000				
Coefficients	**Estimate**	**Std. Error**	**t value**	**Pr (> | t|)**
Baseline	"0.5118"	"83.2258"	"0.0061"	"0.9951"
Seasonality	"0.0385"	"83.2227"	"0.0005"	"0.9996"
Price (conv. unit)	"151.6345"	"81.992"	"1.8494"	"0.0665"
Distribution, Flyers (weighted)	"188.9552"	"81.5968"	"2.3157"	"0.022"
Distribution, Flyers/Display (weighted)	"84.2241"	"82.8378"	"1.0167"	"0.311"
Shopping days	"66.0909"	"83.0169"	"0.7961"	"0.4273"
Covid19 (Lockdown 1)	"0.6246"	"83.2256"	"0.0075"	"0.994"
Covid19 (Lockdown 2)	"1.6472"	"83.2251"	"0.0198"	"0.9842"
TV (GRPs) log adstock 96	"26.6719"	"83.187"	"0.3206"	"0.749"
Display (Imps) log adstock50	"37.6567"	"82.9958"	"0.4537"	"0.6507"
ATV (Spend) log adstock50	"13.2199"	"83.1496"	"0.159"	"0.8739"
Online Video (Imps) log adstock50	"34.1353"	"82.9891"	"0.4113"	"0.6814"
Social (Imps) log adstock50	"41.1332"	"82.9038"	"0.4962"	"0.6205"
---	" "	" "	" "	" "
R-Quadrat = "0.7299"	"0.7299"			
Durbin-Watson = "0.4199"	"0.4199"			
Lambda = "1000000"	"1000000"			

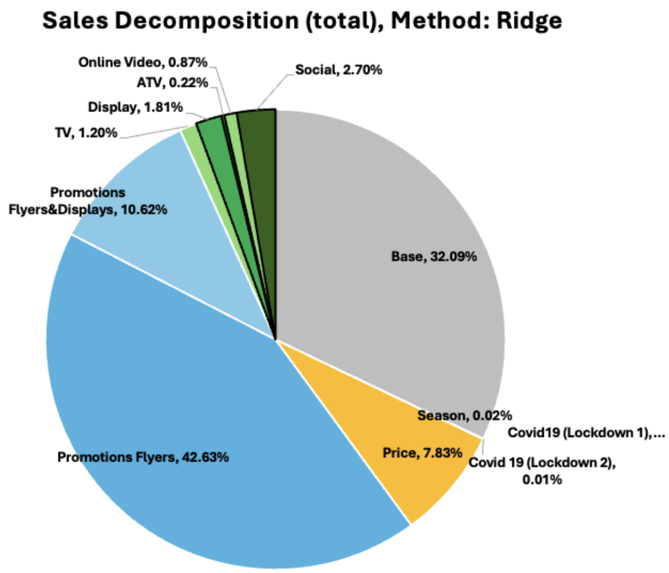
Table 5: Model statistics, ridge regression, with λ = 1,000,000

The R sq. now drops to 0.73 while the individual digital channels are not statistically significant, but all variables are positive in the model (Table 4). The calculated ratio of the coefficients of digital channels in the models are:

Display 285, ATV: 100, OL Video: 258, Social: 311

Thus, the ratio remains comparatively constant, with the coefficients converging further as lambda increases.

Based on the first ridge model, the model output results can be decomposed into individual model drivers and visualized as a sales decomposition (Fig. 5):



Key Driver	Share	Method
Base	32%	Ridge
Season	0%	Ridge
Price	8%	Ridge
Promos	54%	Ridge
TV	1%	Ridge
Display	2%	Ridge
ATV	0%	Ridge
OL video	1%	Ridge
Social	3%	Ridge

Fig. 5: Sales decomposition based on the ridge regression model with λ = 10,000

We notice a clear change in the proportions of the calculated impact factors compared to the sales decomposition obtained from the original OLS model. Thus, the exclusive application of ridge regression techniques results in a "biased" model, when compared to the model output results of "classical" multiple regression.

Therefore, it is not recommended to apply ridge regression directly when conducting Marketing Mix Modeling. This is partially due to the use of non-significant variables in the model and also because of the "intervention-effect" of a regulatory factor in the equation for determining the regression coefficients.

b) Two-stage Model

If the outcomes of the ridge regression are directly employed, as discussed, it leads to a misinterpretation of the impact of the independent variables (see model comparisons in Fig. 6 and Fig. 8). However, given the relative stability of the relationship among the variables in the ridge regression model, it is advisable to adopt a sequential approach as follows (despite variations in the regularization factor λ):

I. The procedure entails the computation of a model utilizing classical multiple linear regression, wherein the "granular" variables of interest are consolidated into a collective variable (sum variable). A prerequisite for this consolidation process is the uniformity of units and the substantiated significance of the grouped variables in the initial model.
II. Subsequently, a model is computed using ridge regression techniques, incorporating an appropriate correction factor λ, to appropriately account for the effects of the "granular" variable.
III. Following this, the sum variable, which showed statistical significance in the initial model, is partitioned based on the model coefficients of the "granular" variables which are determined through ridge regression modeling.

This approach offers several advantages:

- Statistical significance can be established for the sum variable.
- The distribution of variables aligns with the outcomes of a "legitimate" model.
- The relationships among variables remain stable across various values of λ above a certain threshold.
- The effects of other variables, which were declared statistically significant in the original model, remain unaltered.

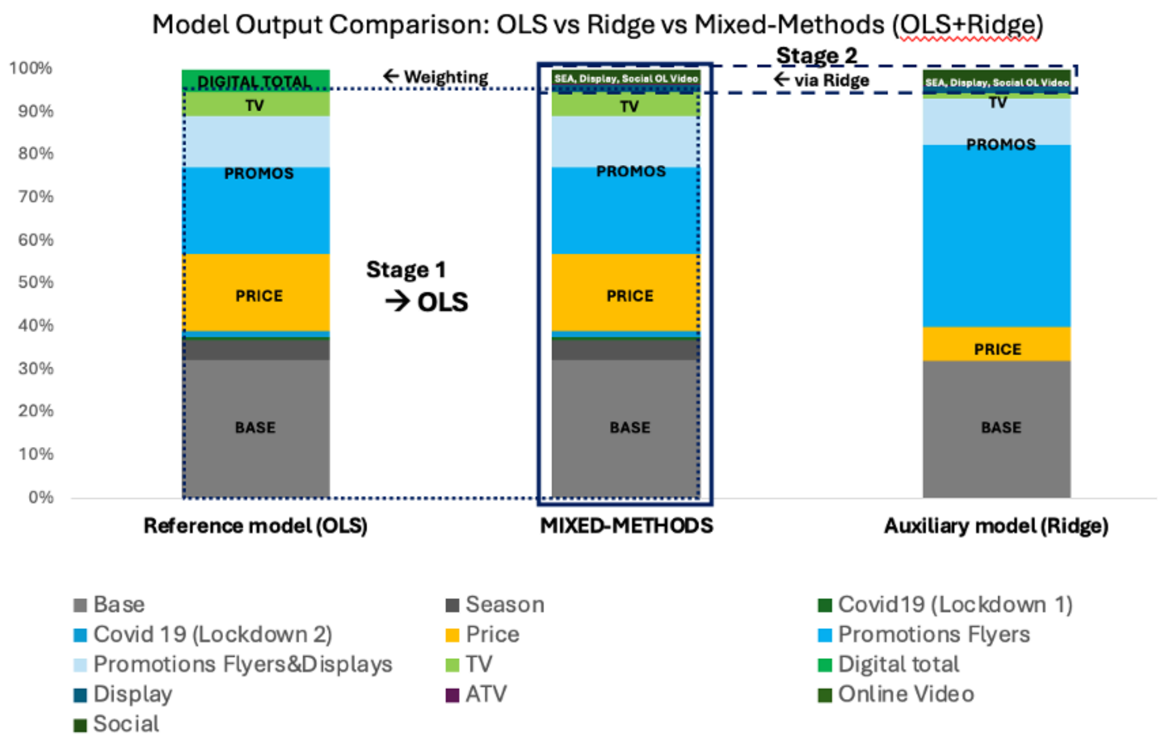
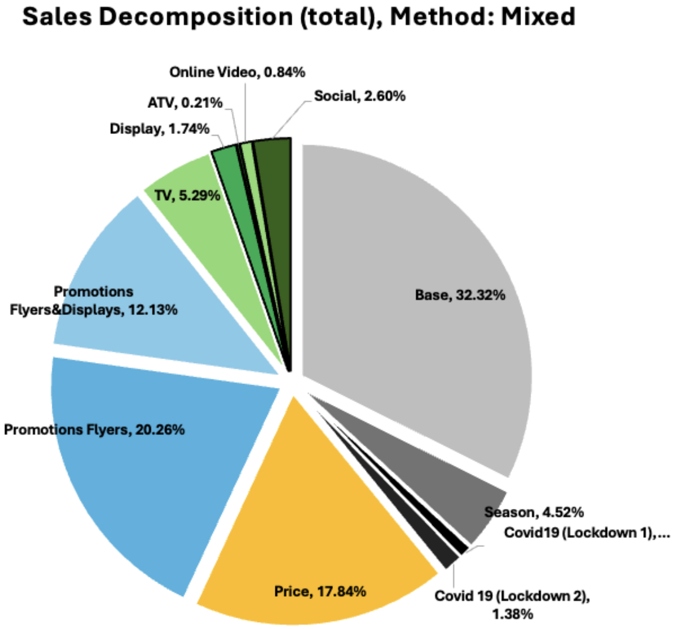


Fig. 6: Sales decompositions of the different models through the modeling process

In this example, the following sales decomposition can be calculated after the final weighting process for the digital channels via the two stage-modeling process (Fig.7).



Key Driver	Share	Method
Base	32%	OLS
Season	5%	OLS
Price	18%	OLS
Promos	32%	OLS
TV	5%	OLS
Display	2%	Ridge
ATV	0%	Ridge
OL video	1%	Ridge
Social	3%	Ridge

Fig. 7: Sales decomposition based on the two-stage modeling process

c) Comparison of Model Outputs

The reference (OLS) and auxiliary (ridge-only) model outputs are compared in Fig. 8. The respective models show that non-digital impact factors such as “Seasonality”, (Fig. 8, (1)), “Price” (Fig. 8, (2)), and “Promotions” (Fig. 8, (3)) have different contribution levels depending on the model. The variable “Price” is undervalued in the auxiliary (ridge-only) model, while “Promotions” is overvalued in the same model. We use the originally calculated contributions as a reference and can rely on the individual p-values determined in the OLS model (Table1). Similarly, “TV” (Fig. 8, (4)) is massively undervalued in the auxiliary (ridge-only) model as shown below. On the other hand, the reference (OLS) model cannot provide any information on the individual digital only channels due to relatively low spend levels among the individual channels and associated issues with multicollinearity. Therefore, we can use the calculated contributions of the individual digital channels in the auxiliary model. However, we have to weight the calculated contributions by using the modeled digital total buckets as modeled in the reference model. This “Two-stage / Mixed-Methods” approach ensures that impact factors are not under-/overvalued in the modeling process while producing overall statistically robust models. This is essential for the subsequent budget allocation stage.

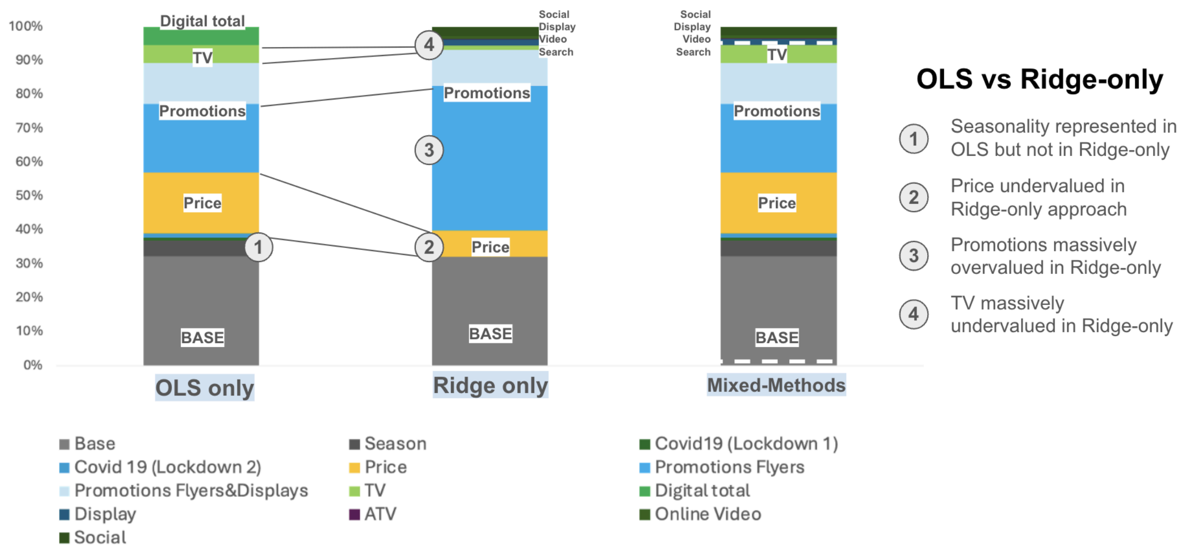


Fig. 8: Model output comparison of sales decompositions

d) Cross-case results (Database analysis)

In recent years, the procedure described above has been frequently conducted by bynd Consulting to detect the effects of individual digital channels on sales outcomes. There were very few cases where not all of the considered digital measures could be detected even when using this approach. This primarily pertained to instances where a single medium, such as online audio, was utilized with minimal advertising pressure. However, the effects of digital media with higher advertising pressures, such as display, online video (OL Video), social, and search advertising (SEA), could always be detected without any difficulty. When performing model updates for extended data periods with similar further media usage, similar results regarding the effects compared to the initial analysis were obtained. Thus, ridge regression as applied in the “Mixed-Methods MMM” approach can be considered a relatively stable technique that generates comparable results under similar conditions.

Collectively, the analysis of these cases generate insights into the average effect per contact and medium as determined based on ridge regression methods. The comparison of model coefficients serves as the basis for this assessment, with the coefficient for OL Video set to an index value of 100 per case. Given that the effectiveness of media diminishes with increasing advertising pressure due to diminishing marginal returns, these average effects per contact should be interpreted as the effect per contact at an equal number of contacts. The effects were examined across different brands and KPIs, including sales and brand parameters. Despite the case-by-case normalization, the averages across digital media remain interpretable and comparable (Fig.9).

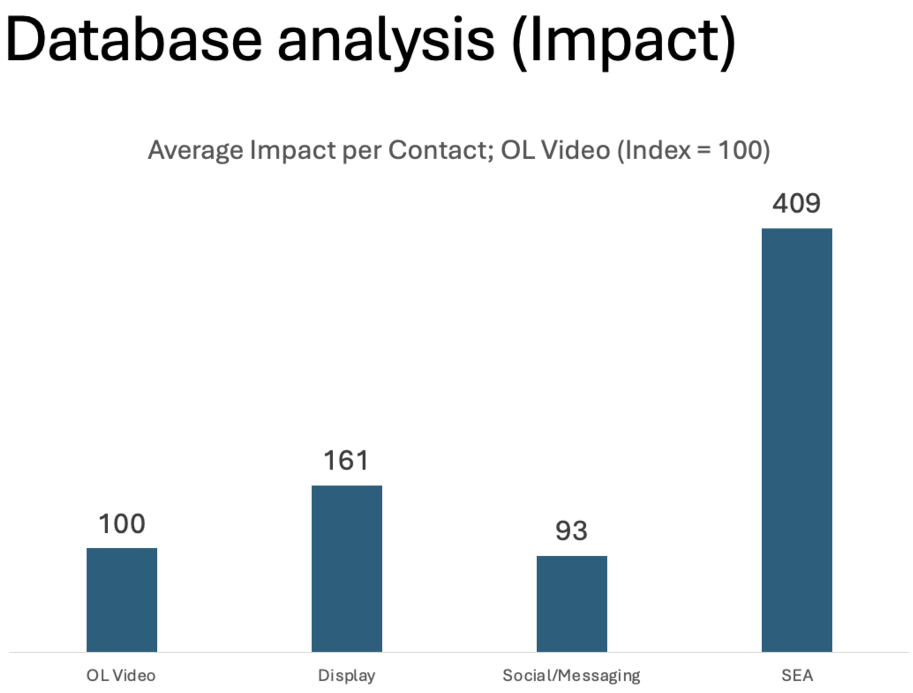


Fig. 9: Average impact per contact, n = 33 cases

It will come as no surprise that the effect per contact is strongest for Search Engine Advertising (SEA). This can be attributed to the fact that specific searches are often conducted towards the end of a "customer journey". In addition, we were able to demonstrate in many cases that searches (for which the trajectory of Google search queries was used as an indicator) are at least partially triggered by the use of media - both classic and digital. While the effect per contact for social and messaging platforms is comparatively low in direct comparison, the costs for a social contact typically fall below those of contacts made through other digital media considered here.

Regarding Display advertising, it is noteworthy that the greatest variation in the effect per contact was observed in this domain. This variation may be attributed to the frequent communication of specific offers in Display advertising, resulting in a distinct difference in effect compared to the effect of the “creations” communicated.

A slightly different perspective emerges when examining the return on investment (ROI/ROAS) across the available cases in which sales (volume or revenue) were modeled. This subset constitutes a portion of the cases analyzed above. Here, the ROAS for each case was indexed, with the ROAS for Online Video (OL Video) assigned an index value of 100. Once again, SEA stands out as the medium with the highest efficiency; however, there is a comparatively high efficient ROAS result for Social/Messaging as well. It is important to note that the ROAS is not independent of usage; factors such as the effect per contact, advertising pressure (resulting in lower ROAS levels with higher advertising pressures due to diminishing marginal returns), and the cost per contact also play a role (Fig.10).

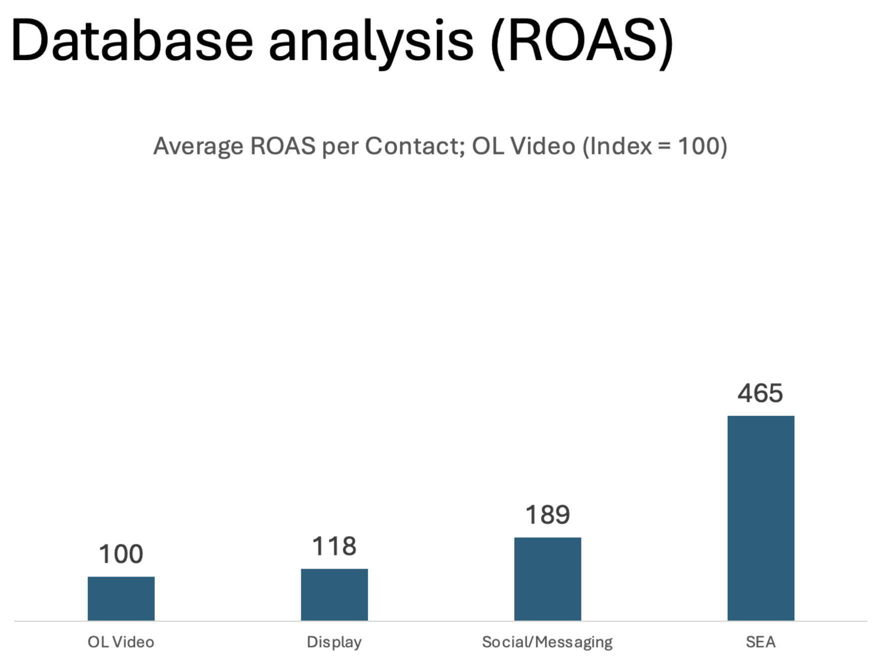


Fig. 10: ROAS, n = 17 cases

e) General recommendations

Based on past experience, we can derive following general recommendations:

- Choosing the correct Lambda factor

It has been demonstrated that the relationship (ratio) of the coefficients for the variables under consideration remains approximately stable above a certain value. We recommend using the smallest value for lambda, if the coefficients of the media variables have a positive sign and are approximately stable to each other.

- Maintaining carry-over factors

In our experience, the coefficients in a ridge regression approach respond more strongly to changes in the carryover than in "classical" linear regression. In order not to skew the results, we recommend using equal carryover rates for all variables to be split by the utilized ridge regression. Different carryover rates should only be used if there are concrete indications or empirical values for different effects.

- Variable set

In order to ensure maximum comparability, the model created in the ridge regression split should contain the same set of variables, except for the variables to be involved in the split process. A modification of the variable set should only be made in exceptional cases.

- Same scale or unit of variables

The variables to be split by applying ridge regression methods should always be represented by the same unit of measurement. Unlike "classical" linear regression, in which the coefficients adjust for different units of the independent variable, in ridge regression the result can change when using different units.

4. Summary

When applying ridge regression, it can be used to quantify "granular" measures more accurately by avoiding both overestimation and underestimation of their respective effects, provided a two-step approach is adopted. The model outcomes hinge upon a regression process, enabling the assessment of the measures' significance through evidence obtained in the initial model.

The combination of classical multiple linear regression in the first stage and ridge regression methods in the second stage facilitates a realistic impact evaluation of all variables considered in the models. Furthermore, both stages generate regression-based outcomes, thereby determining optimal weights for modeled variables is crucial in the modeling process.

5. Conclusion

Although the proposed modeling approach performs well in practice and has proven effective in many cases, there are still several open tasks for future research. These tasks include:

- Specifying a specific error probability for the coefficients within the ridge regression.
- Estimating an optimum lambda based on the data structures rather than through "trial and error".
- Defining requirements for variables to be partitioned within the ridge regression modeling process (number of data points, advertising pressure, etc.).
- Calculating a meaningful maximum number of variables that can be partitioned by the ridge regression.

References

Berk, R. A. (2020). *Statistical learning from a regression perspective*. Springer Nature.

Bortz, J. & Schuster, C. (2016). *Statistics for Humanities and Social Sciences: Limited Special Edition.* Knight.

Broadbent, S. (1979). *One Way TV Advertisements Work*. Journal of the Market Research Society, 23(3), S. 139-166.

Durbin, J. & Watson, G. S. (1971). *Testing for serial correlation in least squares regression*. III. Biometrika, 58, S. 1-19.

Jin, Y., Wang, Y., Sun, Y., Chan, D., & Koehler, J. (2017). *Bayesian Methods for Media Mix Modeling with Carryover and Shape Effects*. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/46001.pdf

O'brien, R. M. (2007). *A caution regarding rules of thumb for variance inflation factors - quality & quantity*. SpringerLink. https://link.springer.com/article/10.1007/s11135-006-9018-6

Pearson, K. (1908). *On the generalized probable error in multiple normal correlation*. Biometrika, 6, S. 59-68.

Preuss, A. (2019). *Econometric Modelling: Gaining insights from data.* In: Markenartikel, 3/2019, pp. 20-23.

Zhou, G., Sentana, L., Skokan, I., & Prada, A. (2021). *Marketing Mix Modelling using multi-objective Hyperparameter optimization*. Technical Disclosure Commons. https://www.tdcommons.org/dpubs_series/4627/