Data Mining

Regression

Santiago Alonso-Díaz

Tecnólogico de Monterrey
EGADE, Business School

Photo: Dalle2

Regression is the go to method in business, economics, finance, and analytics in general.

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise **y_2-y_1** differences. - (Same, but it predicts the *signed rank* of **y_2-y_1**.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…[B])[A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())` As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

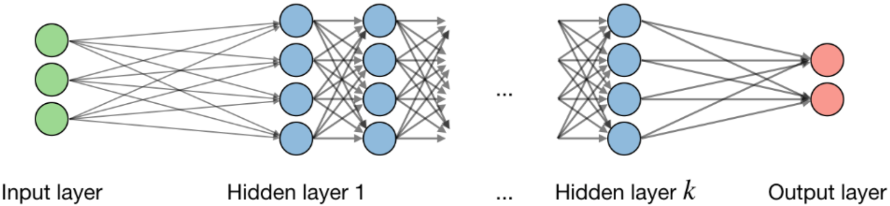
List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G₂, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Figure: Many popular analysis are regressions

Deep learning, AI, and linear sums



By noting i the i^{th} layer of the network and j the j^{th} hidden unit of the layer, we have:

$$\boxed{z_j^{[i]} = {w_j^{[i]}}^T x + b_j^{[i]}}$$

where we note w, b, z the weight, bias and output respectively.

Figure: Source: Shervine Amidi

AI intuitions and regressions (Chollet creator of Keras)

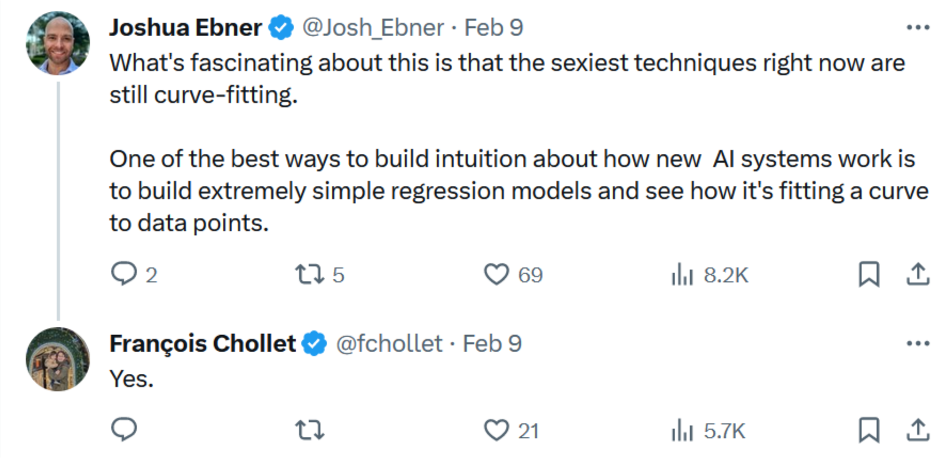


Figure: 2024 exchange in X

Regression still predicts well in business (Schmitt, 2023)

Table: Credit risk prediction

Method	**Out-of-Sample Performance**			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.712	0.671	0.653	0.623
Random Forest	0.773	0.711	0.688	0.572
Gradient Boosting Machine	**0.774**	0.712	0.691	0.572
Deep Learning + ReLU	0.760	0.700	0.646	0.592
Deep Learning + Maxout	0.762	0.703	0.687	0.599

Regression still predicts well in business (Schmitt, 2023)

Table: Insurance claims predictions

Method	**Out-of-Sample Performance**			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.629	0.594	0.586	0.667
Random Forest	0.636	0.598	0.584	0.667
Gradient Boosting Machine	**0.640**	0.602	0.588	0.664
Deep Learning + ReLU	0.628	0.597	0.540	0.670
Deep Learning + Maxout	0.633	0.597	0.534	0.669

Regression still predicts well in business (Schmitt, 2023)

Table: Marketing and sales predictions

Method	**Out-of-Sample Performance**			
	AUC	**Accuracy**	**F-score**	**Logloss**
Logistig Regression	0.918	0.839	0.845	0.377
Random Forest	0.940	0.879	0.888	0.320
Gradient Boosting Machine	**0.940**	0.878	0.886	0.299
Deep Learning + ReLU	0.930	0.861	0.877	0.328
Deep Learning + Maxout	0.930	0.857	0.865	0.336

Table of contents

Linear Regression (James et al., 2023)

Advertising should work like this

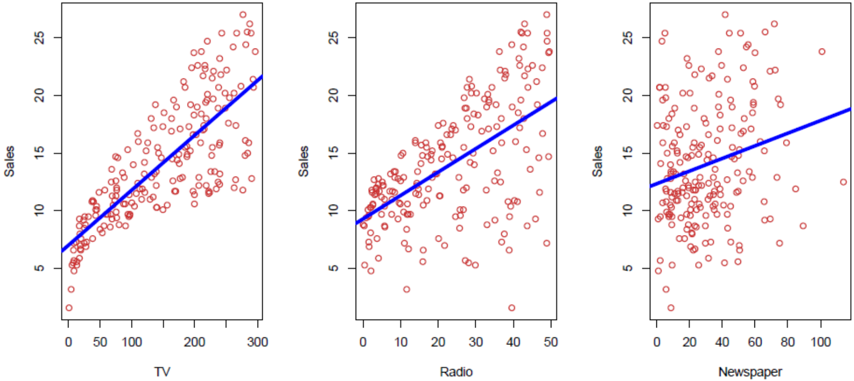


Figure: James et al., 2023

Advertising should work like this

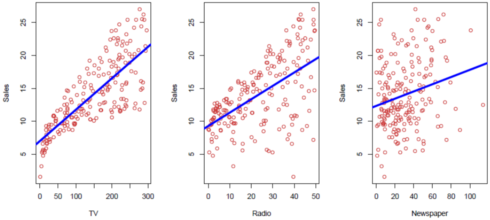


Figure: James et al., 2023

This is the simple regression line:

- $sales = \beta_0 + \beta_1 TV$
- $sales = \beta_0 + \beta_1 Radio$
- $sales = \beta_0 + \beta_1 Newspaper$

How can we estimate the weights?

Loss functions (distance between data and model):

$$min[L(data, model)]$$

In linear regression L is usually the mean squared error:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

with:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}X_i + unobserved_i + \epsilon$$

A picture is worth ...

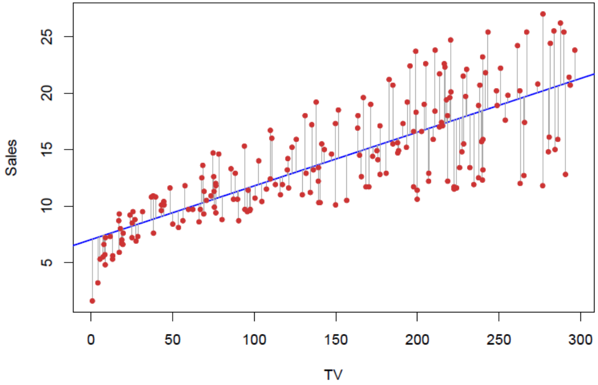


Figure: Line that minimizes MSE (James et al., 2023)

A table is worth ...

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Figure: DV: Sales. Parameters of the line that minimizes MSE (James et al., 2023)

- What is each coefficient? Marginal effect
- What is Std. Error? Coefficient uncertainty
- What is the null hypothesis statistic? t
- What is p-value? p(Effect|Null)

A table is worth ...

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	7.0325	0.4578	15.36	< 0.0001
`TV`	0.0475	0.0027	17.67	< 0.0001

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

Figure: DV: Sales. Parameters and fit info (James et al., 2023)

- What is residual standard error?
- What is R^2?
- What is F-Statistic?

What about more variables?

Similar setup with more variables

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} ... + \beta_n x_{i,n} + unobserved_i + \epsilon$$

For instance,

$$sales_i = \beta_0 + \beta_1 TV_i + \beta_2 Radio_i + \beta_3 Newspapers_i + unobserved_i + \epsilon$$

Lines to (hyper)planes

No longer we can visualize it with a line. But we can still use the same loss function.

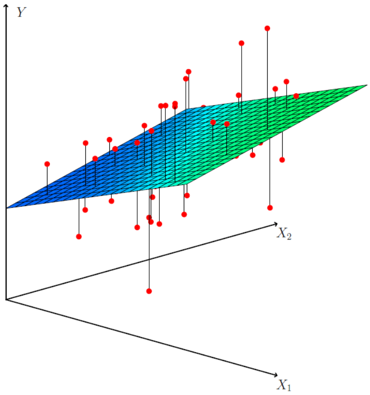


Figure: 4D or more is hard/impossible (James et al., 2023)

We rely more on tables when multiple variables

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

Figure: DV: Sales. Note how newspaper is no longer significant when including more variables (James et al., 2023)

Remember the underlying structure

The non-significant effect of newspapers in the previous regression has to be interpreted under this assumption.

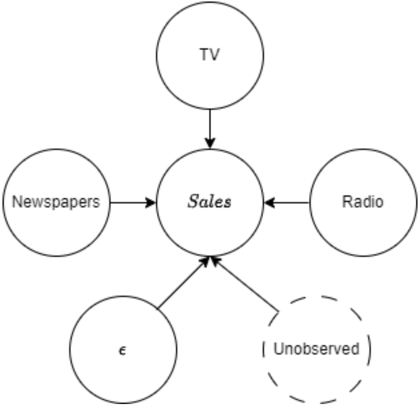


Figure: Relationship assumptions

Remember the underlying structure

But it is not hard to see that there's endogeneity.

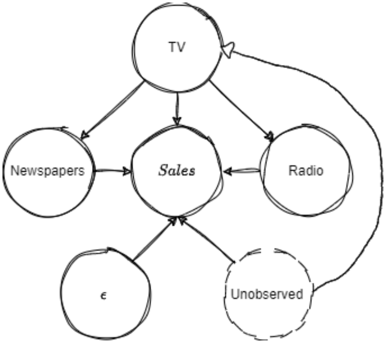


Figure: What if this is the real structure? Or perhaps another? Without newspapers? More on automatic or criteria based variable selection later in the course

Qualitative variables

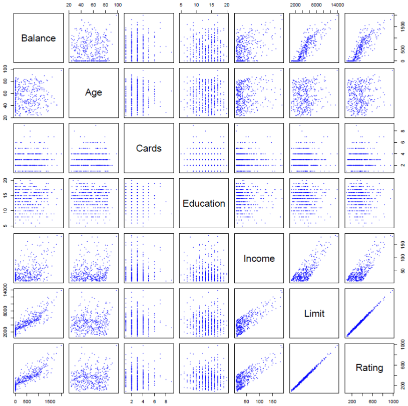


Figure: Credit card data. Not shown here are **qualitative variables**: own (house ownership), student (student status), status (marital status), and region (East, West or South).(James et al., 2023)

Qualitative variables

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	< 0.0001
region[South]	-12.50	56.68	-0.221	0.8260
region[West]	-18.69	65.02	-0.287	0.7740

Figure: DV: Balance USD. Select an arbitrary reference category (e.g. East; won't be in the table). Table estimates are relative to that reference (South and West seem to have lower balances than East; but p>0.05). (James et al., 2023)

Interactions

For instance, what if I multiply TV and Radio expenditures?

$$sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 TVxRadio$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Figure: (James et al., 2023)

Interactions are a change in marginals when the effects are together.

Marketing plan based on linear regression

- Advertising affects sales
- We can estimate the strength of the relationships (βs).
- Let's focus on radio and TV. Newspapers were not significant.
- We can predict sales with the estimated βs
- Exploit interactions

Additional assumptions

- Residuals are normally distributed around zero (no outliers)
- Variance of the residuals is constant (homoscedasticity)
- No relationships between right hand variables (no collinearity)

Let's go to Python: DM_Regression_1.ipynb

Trou Normand: Bayesian Regression

Bayesian Regression

Got to Python: DM_Regression_1.ipynb

Logistic Regression & Classification (James et al., 2023)

Two approaches for credit default

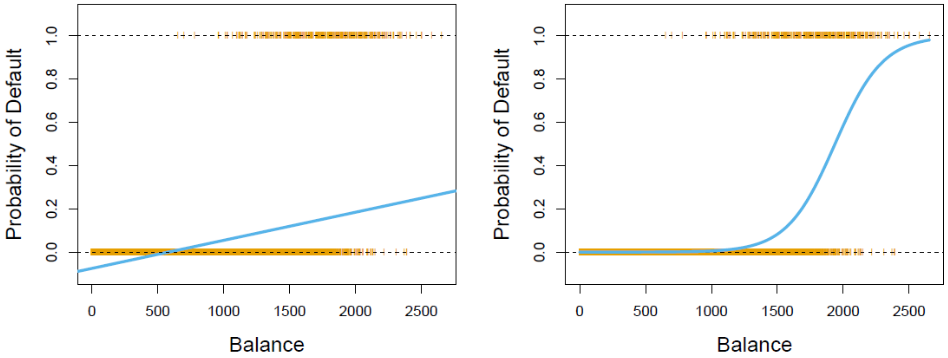


Figure: Both are increasing in balance. The right is an actual probability. Predict default after what balance? (James et al., 2023)

Logistic Regression

What curve to fit? The logistic is a popular sigmoid choice. It is bounded to the interval [0,1] and changes monotonically.

$$p(y = 1|x) = p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

We can even calculate the odds of one outcome over the other. After some algebra:

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0+\beta_1 x}$$

Or in log odds (or logit) is the linear sum of the regression:

$$log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x$$

Logistic regression

How do we estimate the parameters βs? What is the loss function? max. likelihood of the (independent) data given the parameters (MLE).

$$\ell(\beta_0, \beta_1, ..., \beta_n) = \prod_{i:y_i=1} p(x_i) \times \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

MLE is usually executed with numerical/algorithmic methods i.e. no mathematical formula.

Logistic regression examples

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	-10.6513	0.3612	-29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Figure: Credit default with continuous variables (James et al., 2023)

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	-3.5041	0.0707	-49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Figure: Credit default with categorical variables (James et al., 2023)

Logistic regression interpretation

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	-3.5041	0.0707	-49.55	<0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

Figure: Credit default with categorical variables (James et al., 2023)

Probability of default given you are a student (seems low):

$$p(y = 1|student = 1) = \frac{e^{-3.5041+0.4049\times 1}}{1 + e^{-3.5041+0.4049\times 1}} = 0.0431$$

Probability of default given you are NOT a student (better for a no student):

$$p(y = 1|student = 0) = \frac{e^{-3.5041+0.4049\times 0}}{1 + e^{-3.5041+0.4049\times 0}} = 0.0292$$

Odds of default to no default, given a student (low):

$$odds = p(y = 1|student = 1)/p(y = 0|student = 1) = e^{-3.5041+0.4049\times 1} = 0.0451$$

Multiple logistic regression

The same logistic function because it is a "squeezer" of any input and places it in the range [0,1]. Now $X = (x_1, x_2, ..., x_n)$

$$p(y = 1|X) = p(X) = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1 + e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}$$

Multiple logistic example

Now student is negative! Conditioned on the other variables, the effect flipped due to confounding in the previous regression.

	Coefficient	Std. error	z-statistic	p-value
`Intercept`	−10.8690	0.4923	−22.08	<0.0001
`balance`	0.0057	0.0002	24.74	<0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	−0.6468	0.2362	−2.74	0.0062

Figure: (James et al., 2023)

In non-linear regressions, the coefficients are not necessarily marginal effects. This means that it is better to set values for the other variables before concluding anything.

Multiple logistic example

	Coefficient	Std. error	z-statistic	p-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Figure: (James et al., 2023)

Probability of default of X: (student = 1, balance = \$1.500, income = \$40K USD)

$$p(y = 1|X) = \frac{e^{-10.869+0.0057\times 1500+0.0030\times 40-0.6468\times 1}}{1+e^{-10.869+0.0057\times 1500+0.0030\times 40-0.6468\times 1}} = 0.058$$

Probability of default of a non-student with the same balance and income:

$$p(y = 1|X) = \frac{e^{-10.869+0.0057\times 1500+0.0030\times 40-0.6468\times 0}}{1+e^{-10.869+0.0057\times 1500+0.0030\times 40-0.6468\times 0}} = 0.105$$

What about more classes?

The probability of being consumer type 1, 2, 3, 4? The probability of picking flavor A, B, C, D?

We use multinomial logistic regression. But let's leave that for another course. Know that it exists.

Still, let's see other classification techniques that also work with many classes in Python DM_Regression_1.ipynb

References

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.

Schmitt, M. (2023). Deep learning in business analytics: A clash of expectations and reality. *International Journal of Information Management Data Insights*, *3*(1), 100146.