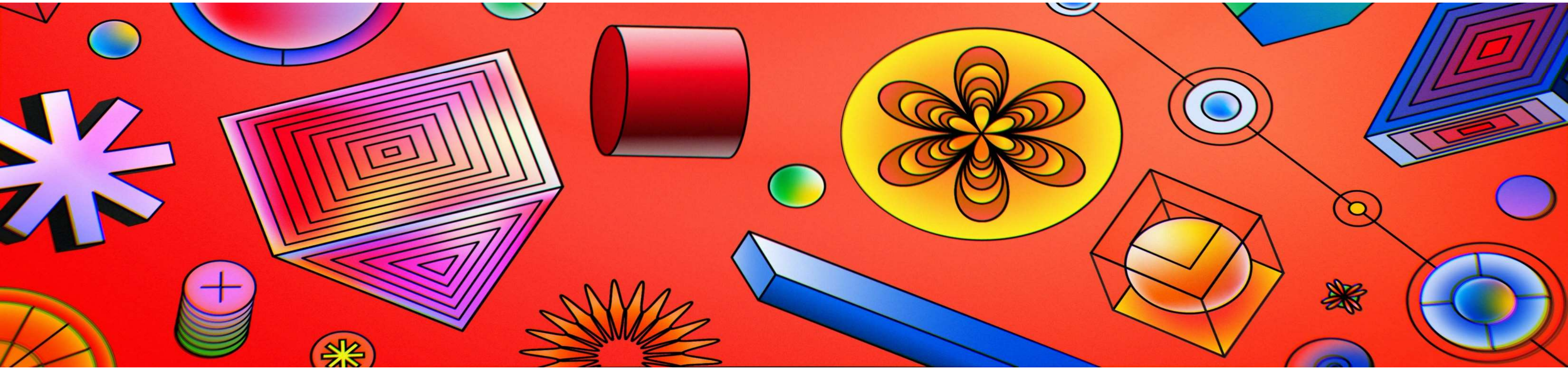
Fall 2023

BIF524/CSC463 Data Mining

Linear Regression

Eileen Marie Hanna, *PhD*

21/09/2023



Coefficients estimates

- Standard errors can also be used to perform **hypothesis testing** on coefficients.

- The most common hypothesis test involves testing the null vs alternative hypotheses:

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$
$$H_a : \text{There is some relationship between } X \text{ and } Y$$

- Mathematically:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Coefficients estimates

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

- To test the null hypothesis, we **need to determine whether $\boldsymbol{\widehat{\beta}_1}$** (our estimate) **is sufficiently far from zero** so that we can be confident that it is non-zero.

- How far is enough?
  - depends on how accurate is our estimate $\hat{\beta}_1$.
    - If $SE(\hat{\beta}_1)$ is small, even relatively small values of $\hat{\beta}_1$ can provide strong evidence that it is non-zero.
    - If it is large, then $\hat{\beta}_1$ must be very large in absolute value so we can reject the null hypothesis.

Coefficients estimates

- In practice, we compute the t-statistic, given by:

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

- We can see that it measures how far $\hat{\beta}_1$ deviates from zero.

  - If no relationship between X and Y -> t-distribution with $n-2$ degrees of freedom.
  - The distribution has a bell shape and for values approx. ≥ 30 -> similar shape to the normal distribution.

H_0 : There is no relationship between X and Y

Coefficients estimates

- It is thus simple to calculate the probability of observing any number equal to or larger than $|t|$, when $\beta_1 = 0$, i.e., the $p - value$.

- small $p - value$ -> a low probability to observe a close relationship between the predictor and the response due to chance, in the actual absence of such relationship.

  - small $p - value$ thus means that there is an association between the predictor and the response -> reject H_0

  - How small?
    - typically, with cut-offs of 1% or 5%

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)}$$

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

$Sales \approx \beta_0 + \beta_1 TV$

Coefficients estimates – "Advertising" dataset

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325 β_0	0.4578	15.36	< 0.0001
TV	$\boxed{0.0475}$ β_1	0.0027	17.67	< 0.0001

sales when TV budget is zero

reflects the effect of TV advertising on sales

The t-statistic is relatively high, and it also corresponds to a very low p-value that we can interpret as:

Small p-value for intercept -> reject the hypothesis that $\beta_0 = 0$ -> when no expenditure on TV, sales are not zero.

Small p-value for TV -> reject the hypothesis that $\beta_1 = 0$ -> there is a relationship between TV and sales.

Model accuracy

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

Model accuracy – residual standard error (RSE)

- Having rejected the null hypothesis -> we need to quantify **how well our model fits the data.**

- $\boldsymbol{RSE}$ estimates the standard deviation of $\in$.

  - It is the **average amount that the response will deviate from the true regression line**.

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Model accuracy – residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **e.g., if $RSE = 3.26$ for our advertising data -> how would you interpret this value?**

$$\text{Sales} \approx 7.0325 + 0.0475 \times \text{TV}$$

Model accuracy – residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- e.g., if $RSE = 3.26$ for our advertising data -> any prediction of sales based on TV advertising would still be off by about 3,260 units on average.

- Whether or not this deviation is acceptable **depends on each problem context**.

- e.g., in the advertising data, the mean value of sales over all markets is around 14000 units -> the percentage error is then $3260/14000 = 23\%$.

Model accuracy – residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RSE is viewed as a measure of lack of fit.

If the predictions obtained from the model are very close to the true outcome values -> RSE will be small -> model fits the data well.

Model accuracy – R^2

- **RSE is measured in units of the output values Y -> not always straightforward to interpret.**

- R^2 is an alternative measure of fit.
  - the proportion of variance explained
  - takes values between 0 and 1
  - independent of Y

the amount of variability that is left unexplained after performing the regression

the amount of variability in the response that is explained by the regression

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

initial total variance in response to Y

R^2 measures the proportion of variability in Y that can be explained using X.

Model accuracy – R^2

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

- **if it is close to 1,**
  - it means that a **large proportion of the variability** in the response has been **explained by regression**.

- **if it is close to 0,**
  - the regression **did not explain much** of the variability.
  - This is when the linear regression is wrong or when the inherent error σ^2 is high (or both).

When the application we are considering is far from being approximated with a linear model -> we expect the value to be close to 0.

Correlation

- Another measure of linear relationship between X and Y, given by:

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

Correlation

- Another measure of linear relationship between X and Y, given by:

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

- For linear regression, squared correlation is equivalent to the R^2 statistic – does not apply to multiple regression that will be covered later.

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Correlation does not imply causality!

Multiple linear regression

- **Typically, more than one predictor that influences a certain response.**

- What if we use different simple linear regression for each of the predictors.

  - What about **dependencies** between predictors?!

  - Use a multiple linear regression which considers **multiple predictors**.

  - **Each predictor** will be given a **separate slope coefficient** in a single model.

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

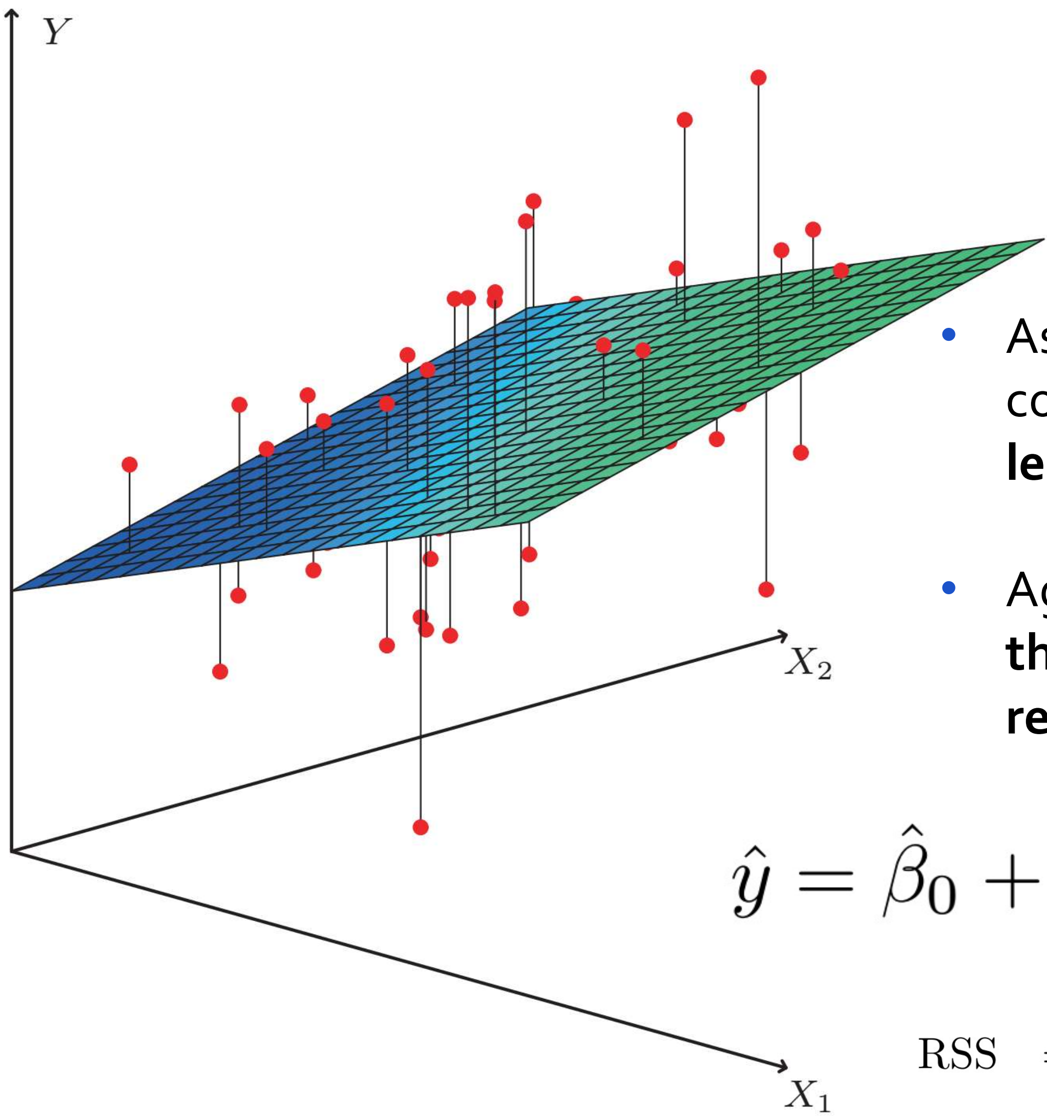
Multiple linear regression

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

Again, β_j is the average effect on Y of a one unit increase in X_j, but if other predictors were fixed!

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon$$

Multiple linear regression – estimating the coefficients



- As in simple linear regression, the coefficients are estimated using the same **least squares approach**.
- Again, we need to choose the coefficients **that minimize the sum of squared residuals (*RSS*)**.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

TV, newspaper, and radio budgets are used to predict sales

	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

What happens if for a given (fixed) amount of TV and newspaper advertising, the client spends extra $\$1000$ on radio advertising?

TV, newspaper, and radio budgets are used to predict sales

Sales $\approx 2.939 + 0.046 \times TV + 0.189 \times radio - 0.001 \times newspaper$

	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

What happens if for a given (fixed) amount of TV and newspaper advertising, the client spends extra $\$1000$ on radio advertising?

We expect an increase of approx. 189 units in sales.

Why not three simple linear regression models?

When TV, newspaper, and radio budgets are used to predict sales

	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

vs. simple linear regression outcomes for each of the predictors

	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

	Coefficient	Std. error	t-statistic	p-value
Intercept	9.312	0.563	16.54	< 0.0001
radio	0.203	0.020	9.92	< 0.0001

	Coefficient	Std. error	t-statistic	p-value
Intercept	12.351	0.621	19.88	< 0.0001
newspaper	0.055	0.017	3.30	0.00115

What can we say about the fact that multiple regression shows no relationship between newspaper and sales, while simple linear regression shows the opposite?

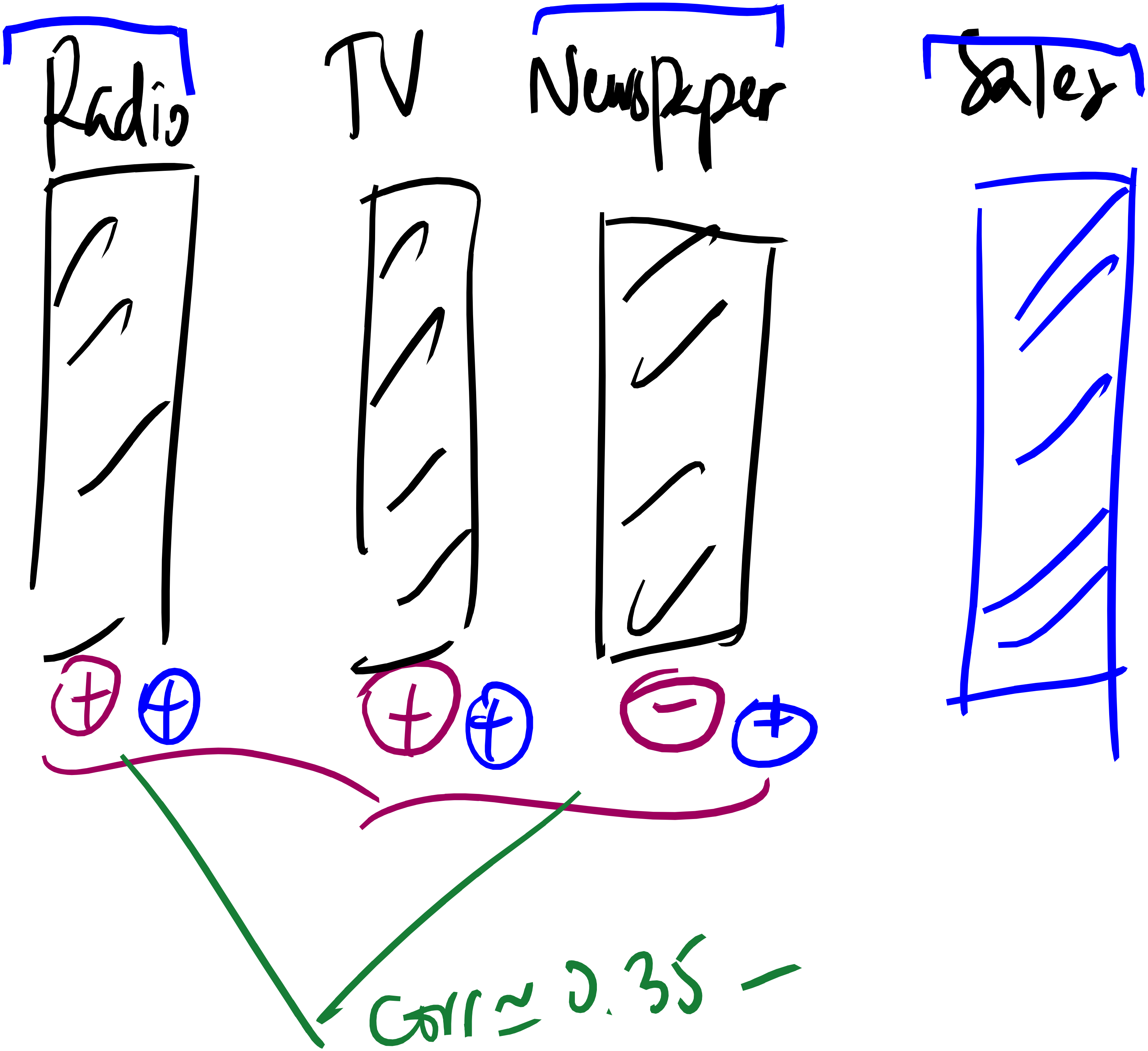
	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

What can we say about the fact that multiple regression shows no relationship between newspaper and sales, while simple linear regression shows the opposite?

	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

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- If we consider the **correlation** values between the predictors and sales,
  - we note **a correlation of 0.3541 between radio and newspaper**
  - tendency **to spend more on newspaper** advertising **when more is also spent on radio** advertising.



Radio
TV
Newspaper
Sales
+
+
+
+
−
+
Corr ≃ 0.35

So again, what are we missing if we only use simple regression?

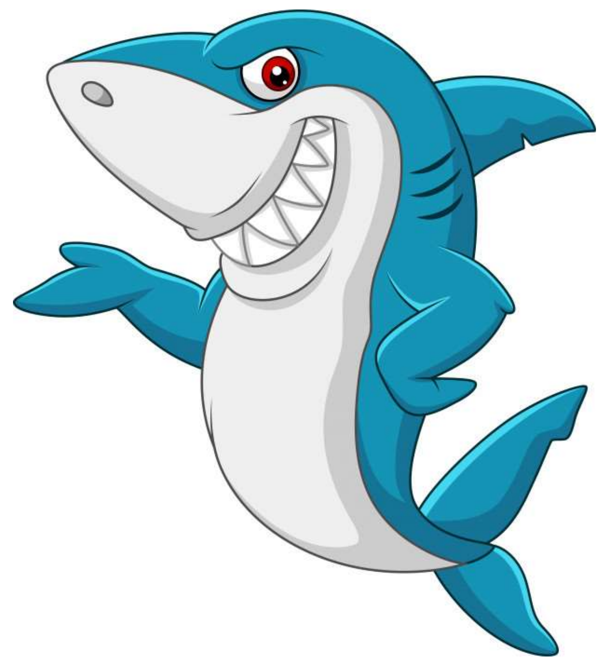
	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	-0.001	0.0059	-0.18	0.8599

- In a **simple linear regression**, which only examines **sales versus newspaper**, **higher values of newspaper** tend to be **associated with higher values of sales**, even though newspaper advertising does not actually affect sales.
- **In a way, newspaper gets “credit” for the effect of radio on sales.**
- A simple regression only considers the effect of newspaper advertising on sales -> increasing newspaper budget leads to sales increase…
- BUT we know based on multiple regression that newspaper advertising does not lead to sales increase!

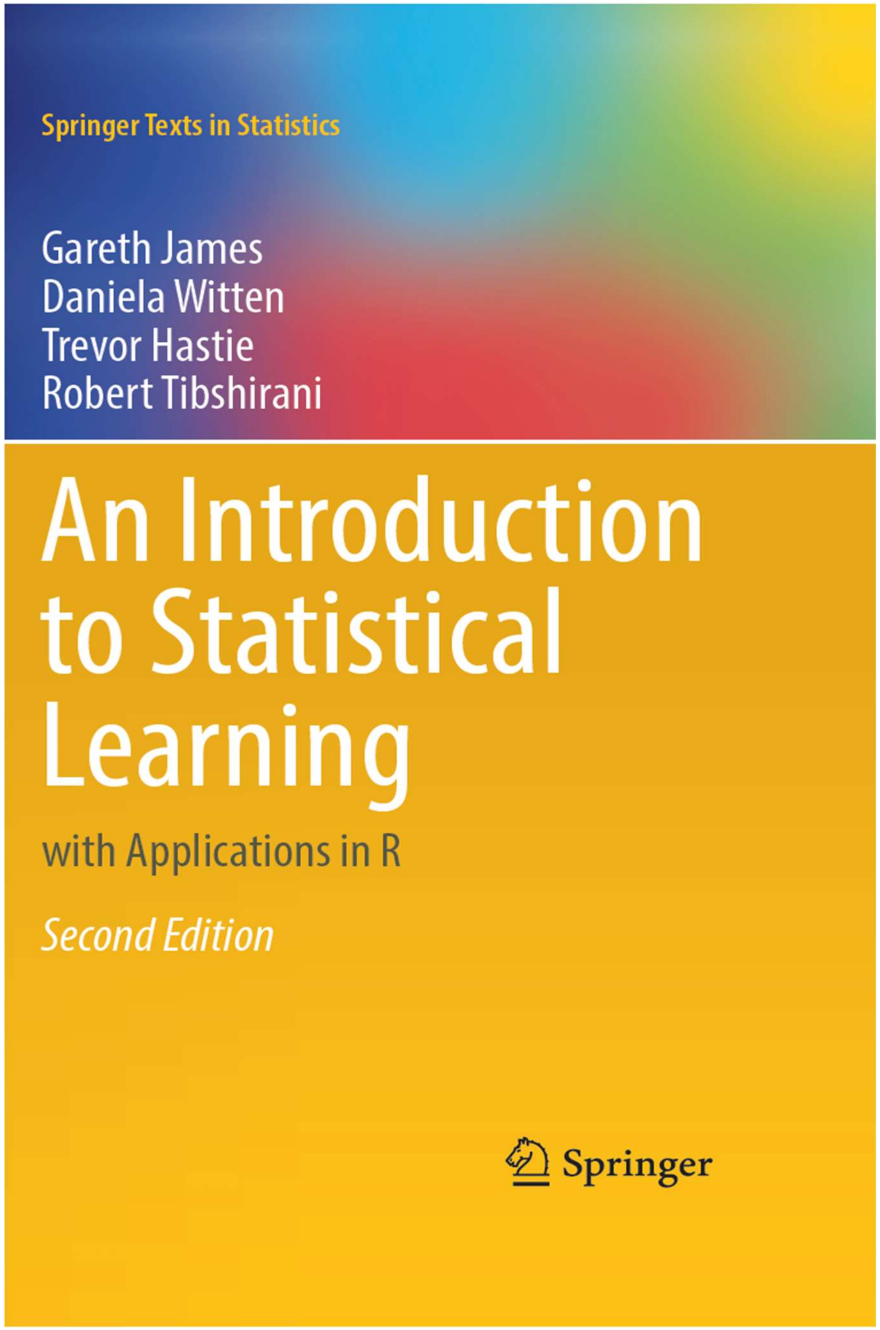
"There is a positive relationship between ice cream sales and shark attacks at a certain beach community".

"Selling ice cream should be banned at this beach community in order to reduce shark attacks".

What if use **Temperature** as predictor in a multiple regression setting?



Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer