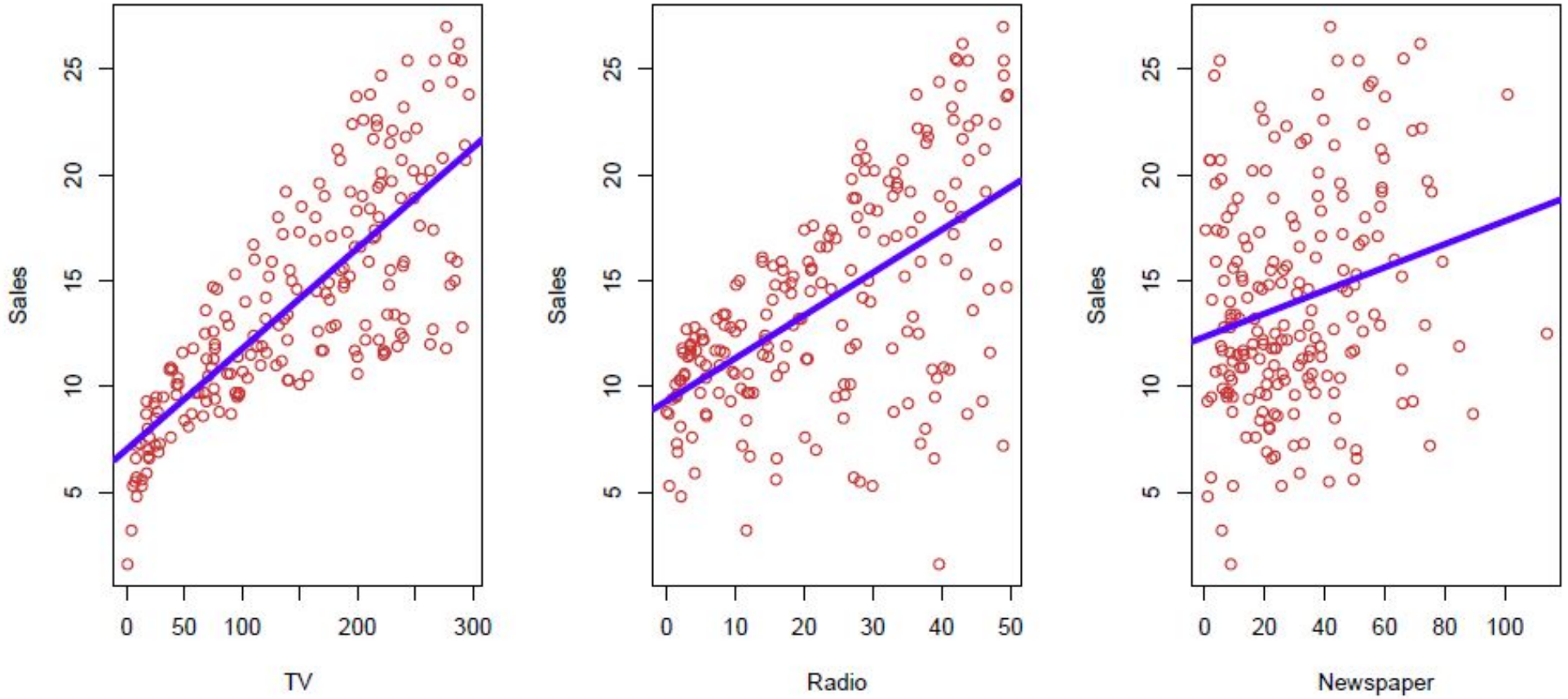
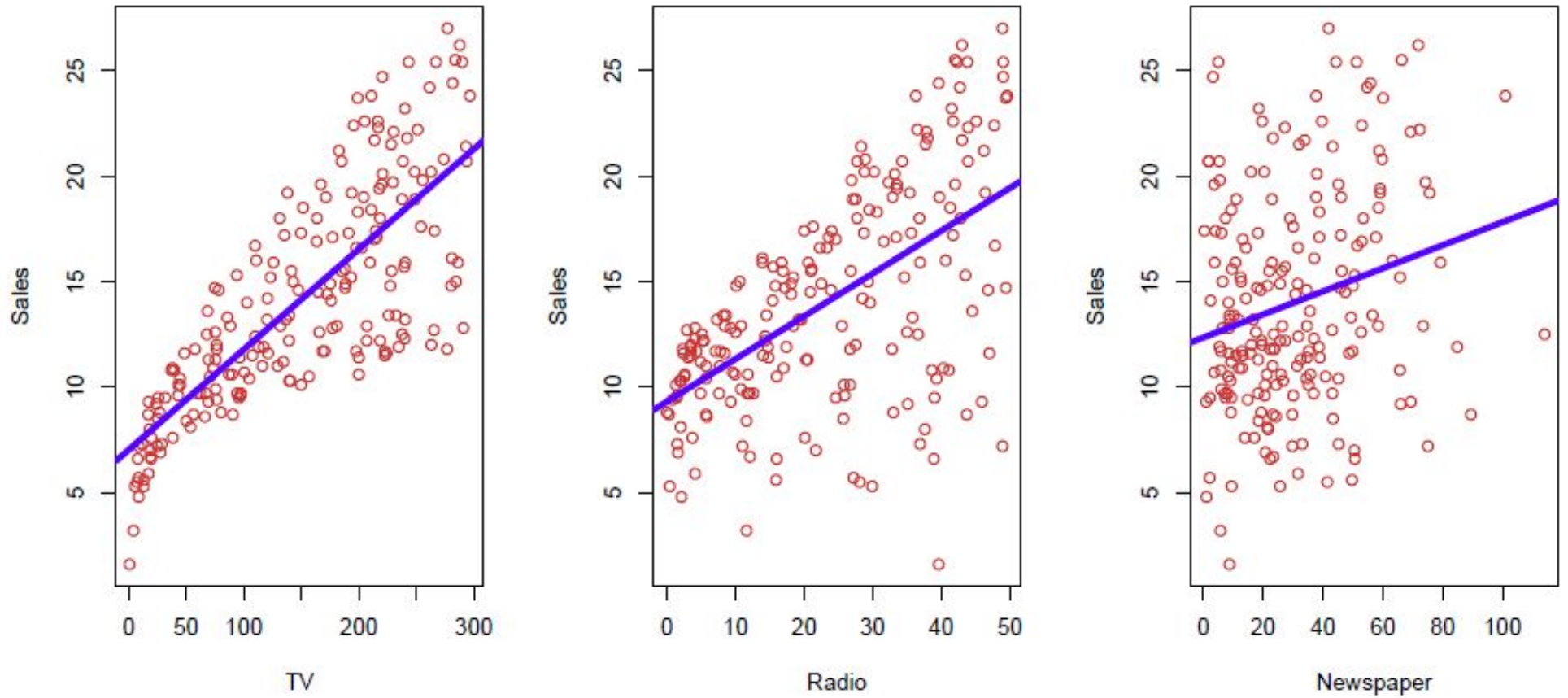
Linear Regresion

Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe

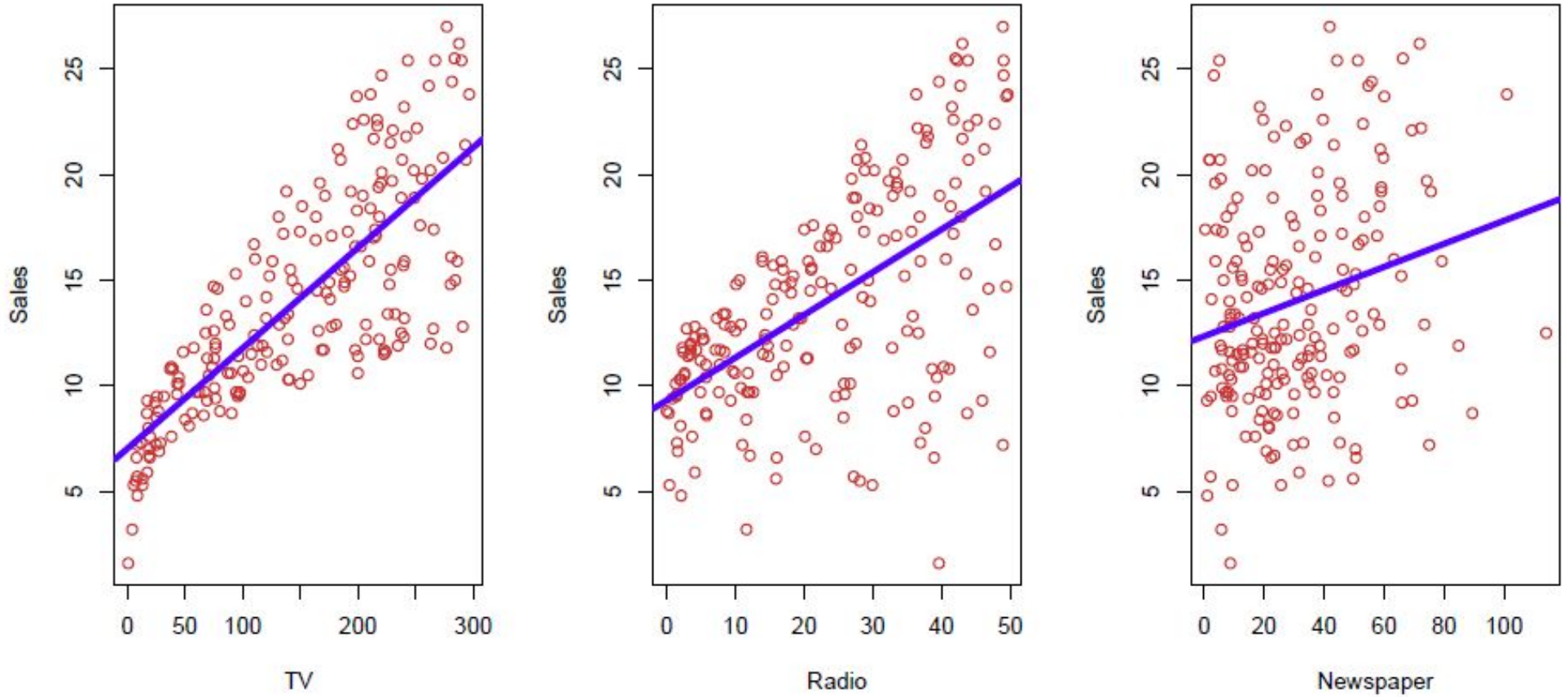
Universidad Católica San Pablo | Departamento de Ciencia de la Computación



Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.



In this setting, the advertising budgets are input variables while sales is an output variable. The input variables are typically denoted using the symbol X, with a subscript to distinguish them. So X_1 might be the TV budget, X_2 the radio budget, and X_3 the newspaper budget. Can we predict Sales using these three? Perhaps we can do better using a model. Now we write our model as $Y = f(X) + \varepsilon$, where ε captures measurement errors and other discrepancies.



- With a good f we can make predictions of Y at new points.
- We can understand which components of X are important in explaining Y, and which are irrelevant.
- Depending on the complexity of f, we may be able to understand how each component of X affects Y.

- **Linear regression** is a simple approach to supervised learning. It assumes that the dependence of Y is linear.

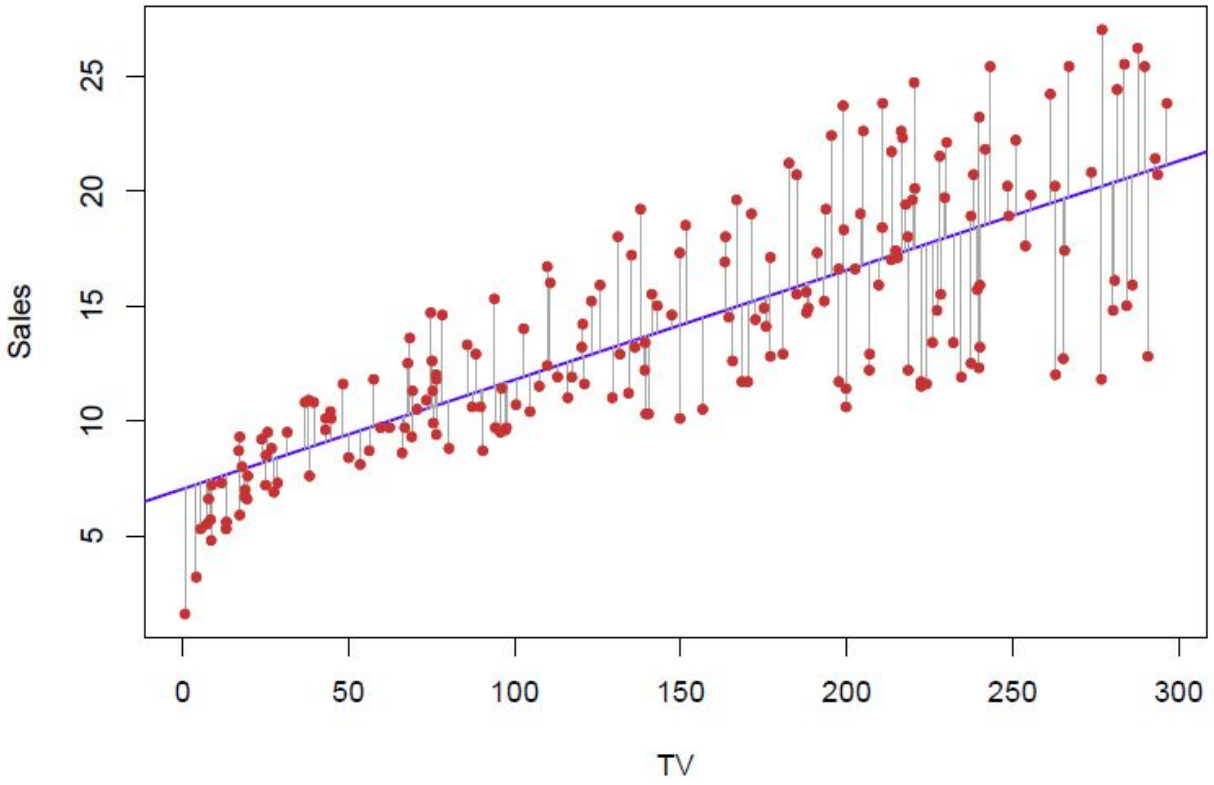
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

- Given some estimates for the model coefficients, we predict Y using

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n$$

- The hat symbol denotes an estimated value.

- We compute the estimated values using the least square method, which consist in minimize the residual sum of squares.



$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

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

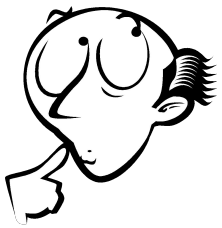
$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

These standard errors can be used to compute confidence intervals. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of

H_0 : There is no relationship between X_1 and Y.
H_a : There is some relationship between X_1 and Y.



Mathematically, this corresponds to testing

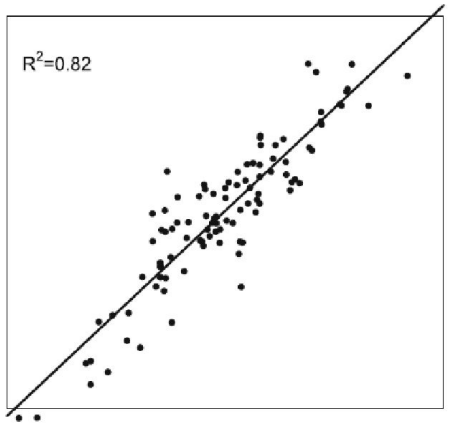
H_0 : β_1=0.
H_a : $\beta_1 \neq 0$.

Since if β_1 = 0 then the model reduces to Y = $\beta_0 + \varepsilon$, and X is not associated with Y .

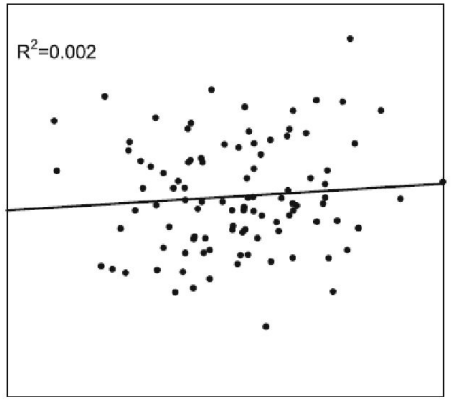
	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

The total variation of *y* is partitioned as:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{SQ_{\text{Total}}}=\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{SQ_{\text{Regression}}}+\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SQ_{\text{Error}}}.$$



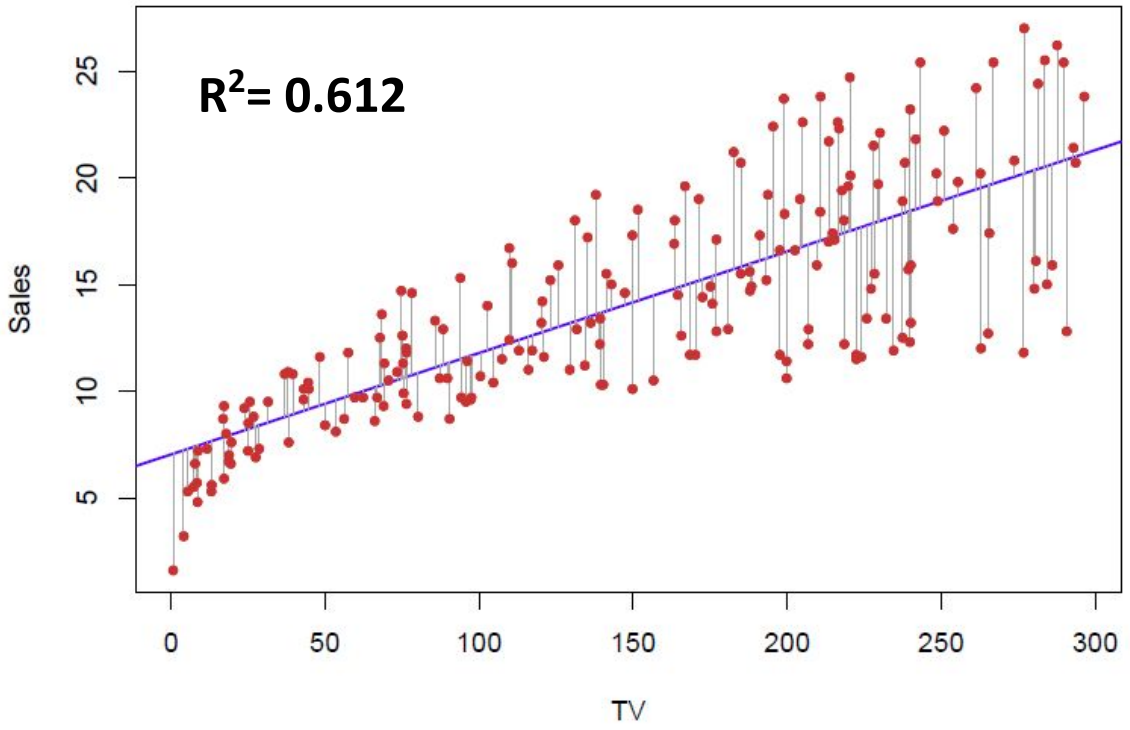
(a) Positive linear relationship



(b) No clear relationship

The criterion for the goodness of fit is given by the **coefficient of determination.**

$$R^2=\frac{SQ_{\text{Regression}}}{SQ_{\text{Total}}}=1-\frac{SQ_{\text{Error}}}{SQ_{\text{Total}}}.$$

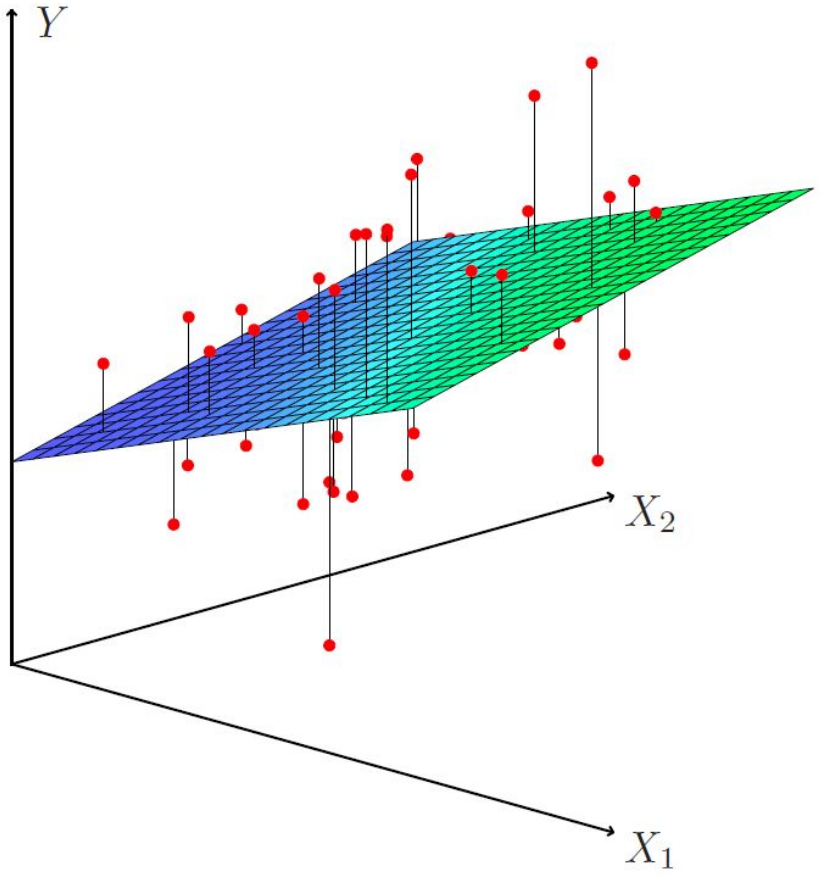


We interpret β_j as the average effect on Y of a one unit increase in X_j , holding all other predictors fixed. In the advertising example, the model becomes

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

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated.
- Correlations amongst predictors cause problems.
- Claims of causality should be avoided for observational data.



Considering all 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

Correlations:

	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

Is There a Relationship Between the Response and Predictors?

In the multiple regression setting with n predictors, we need to ask whether all of the regression coefficients are zero

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0.$$

$$H_a : \text{At least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic. Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, F>1. For the simple regression F-statistic is equal to 312.1 and for multiple linear regression model, we have that the F-statistic is 570.

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Quantity	Value
Residual Standard Error	1.69
R^2	0.897
F-statistic	570

Do all the predictors help to explain Y, or is only a subset of the predictors useful?

It is possible that all of the predictors are associated with the response, but it is more often the case that the response is only associated with a subset of the predictors. The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as variable selection.

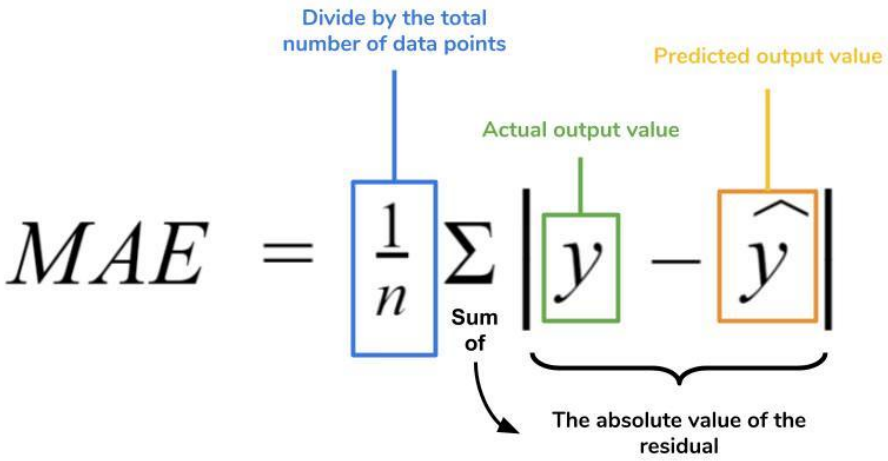
Various statistics can be used to judge the quality of a model. These include Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2.

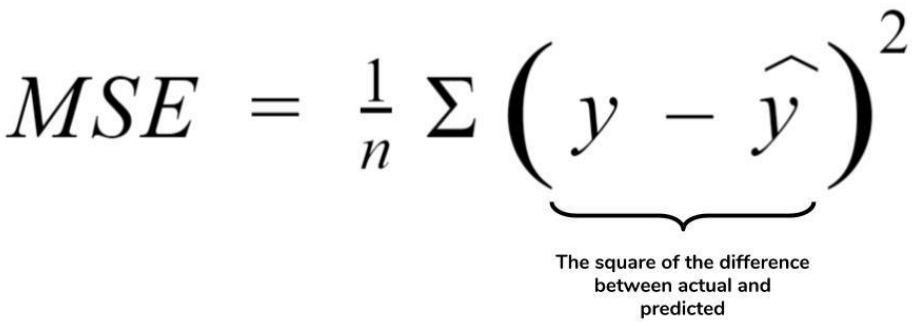
$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

How well does the model fit the data?

- Two of the most common numerical measures of model fit are the RSE and R^2, the fraction of variance explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

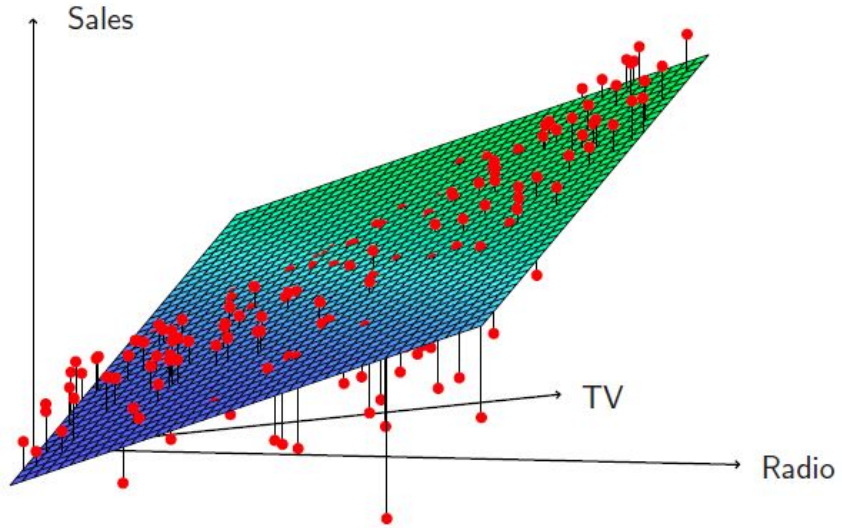
- If our collection of errors are small, it implies that the model that produced them does a good job at predicting our output of interest. Otherwise is a poor estimator. The most used are: Mean Absolute Error (MAE) and Mean Square Error (MSE). **If the data set has outliers, we recommend use MAE instead of MSE**.





Interactions between predictor variables

Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.In marketing, this is known as a synergy effect, and in statistics it is referred to as an **interaction effec**t.



$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

	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

Also, the R^2 for the interaction model is 96.8%, compared to only 89.7% for the model that predicts sales using TV and radio without an interaction term.

Choosing a model can seem intimidating, but a good rule is to start simple and then build your way up. The simplest model is a linear regression, where the outputs are a linearly weighted combination of the inputs. Consider an extension of linear regression called polynomial regression. The general equation for a polynomial is below.

$$Y = \beta_0 + \beta_1 \mathrm{X} + \beta_2 X^2 + \cdots + \beta_n X^n + e.$$