Statistics and Big Data

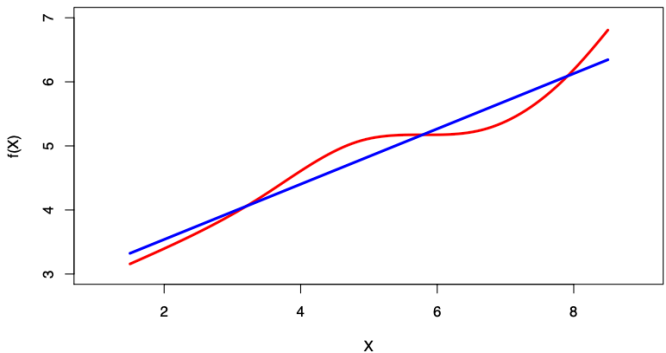
2025-2026

Vincenzo Nardelli

vincenzo.nardelli@unicatt.it

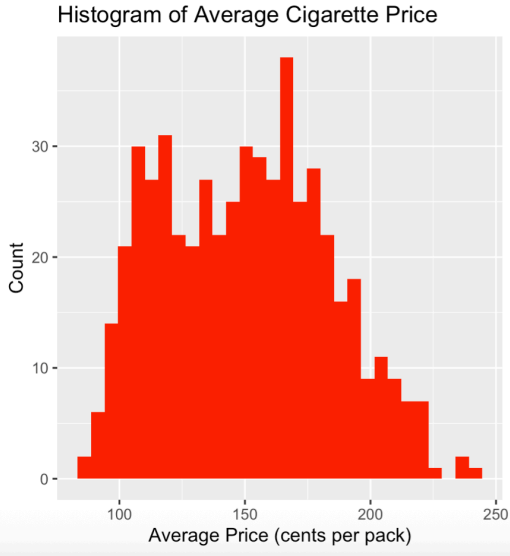
Linear Regression

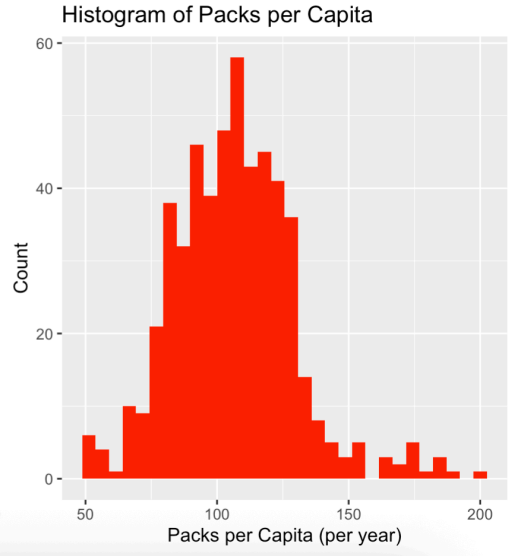
- Linear regression is a simple approach for supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots, X_p$ is linear.



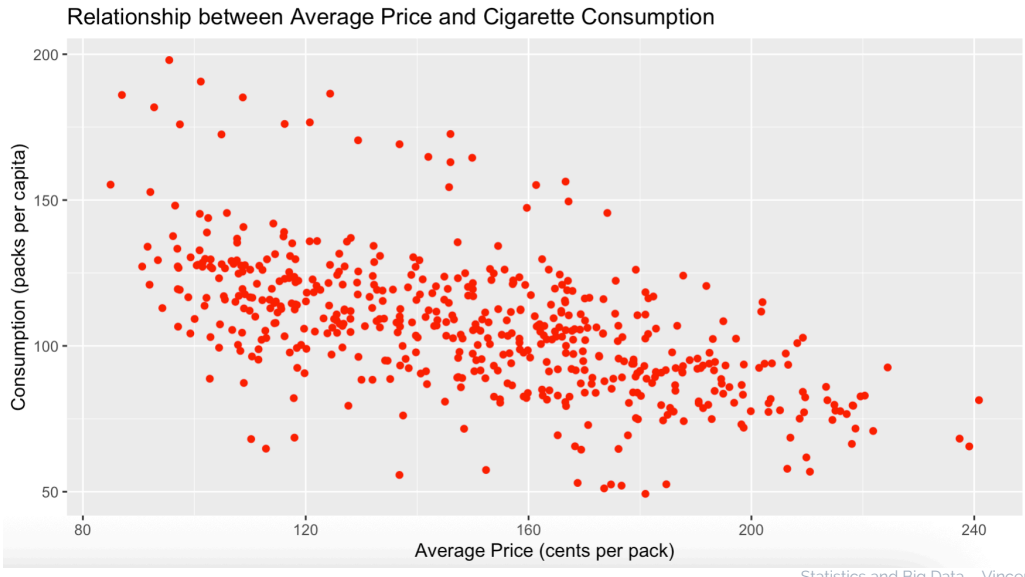
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

Cigarette Data

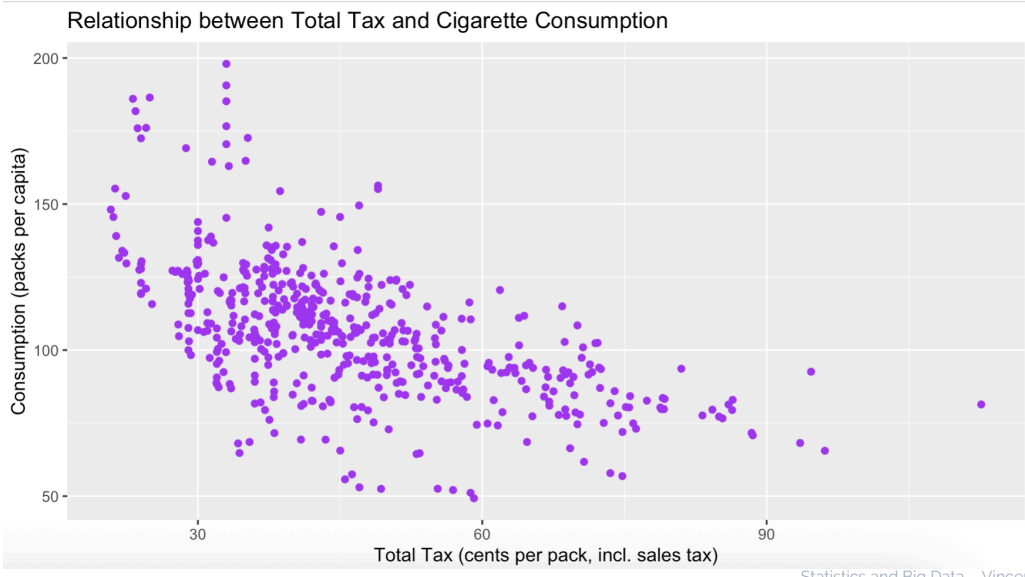




Cigarette Data



Cigarette Data



Linear regression for Cigarette data

Questions we might ask:

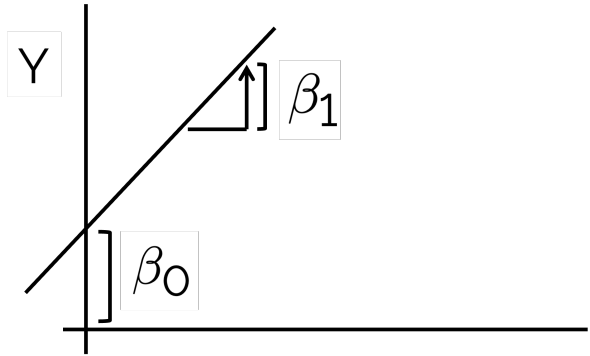
- Is there a relationship between price and consumption?
- Is the relationship linear?
- How strong is the relationship between taxes and price?
- How accurately can we predict future consumption?
- Is there an effect on consumption when taxes increase?

Simple linear regression with a single predictor X

We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and the *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.



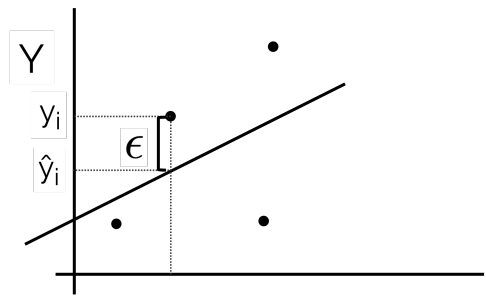
Simple linear regression with a single predictor X

Given some estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ denotes a prediction of Y based on $X = x$. The hat symbol denotes an estimated value.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the i-th value of X. Then $e_i = y_i - \hat{y}_i$ represents the i-th residual.



Parameter estimation via least squares

- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

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

Parameter estimation via least squares

- We define the *residual sum of squares* (RSS) as

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

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

- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

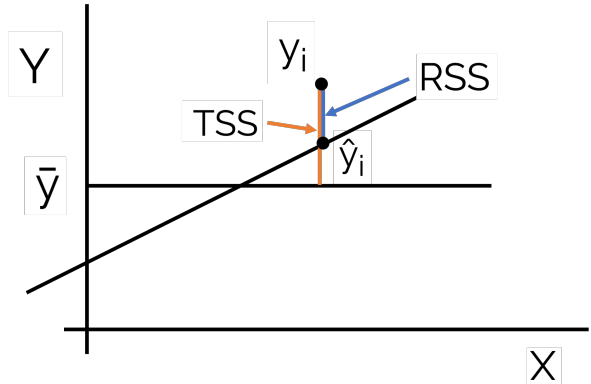
where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Assessing the Overall Accuracy of the Model

- The *R-squared* or fraction of variance explained is

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

where
$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.



LAB Cigarette Consumption

A health agency is studying how price and taxation influence cigarette consumption at the state level. By analyzing historical data, the objective is to understand the effectiveness of policies that increase prices and taxes on tobacco consumption control. This analysis will provide solid foundations for formulating public health recommendations, with the intent to reduce cigarette consumption and health-related risks.

Through the study of these relationships, the health agency intends to evaluate the effect of price and taxation variations to plan strategies that can incentivize a decrease in cigarette consumption.

Variable Description - Cigarette Dataset (Ecdat Package)

The dataset contains information collected at the state level in the United States and includes key variables for analyzing the effect of prices and taxation on cigarette consumption. The main variables are:

- **state**: State where data was collected.
- **year**: Year of data collection.
- **avgprs**: Average price of a cigarette pack (in dollars).
- **packpc**: Cigarette consumption (packs per capita).
- **taxs**: Total taxes on a cigarette pack (in dollars).

LAB Cigarette Consumption: Analysis Questions

- Is there a relationship between the average price of a cigarette pack (`avgprs`) and cigarette consumption (`packpc`)? Is the relationship positive or negative?
- What is the relationship between average price and cigarette consumption? How can we interpret this value to assess the influence of price on consumption?
- How does total taxation (`taxs`) influence cigarette consumption? Analyze this relationship using a scatterplot.
- Through a linear regression model, what would be the impact on cigarette consumption if taxation in a state went from $0.50 to $1.00?

Standard Error

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

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

dove $\sigma^2 = \text{Var}(\epsilon)$.

Confidence Intervals

- These standard errors can be used to compute confidence intervals.
- A 95% confidence interval is defined as an interval that, with 95% probability, contains the true unknown value of the parameter. It has the form:

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}\left(\hat{\beta}_1\right)$$

Confidence Intervals - Continued

- That is, there is approximately a 95% probability that the interval:

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

contains the true value of β_1 (in a scenario where repeated samples like the current one are obtained).

Hypothesis Testing

- Standard errors can also be used to perform hypothesis tests on the coefficients.
- The most common test concerns the null hypothesis:

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

versus the alternative hypothesis:

$$H_A : \text{There is a relationship between X and Y.}$$

- Mathematically, this corresponds to testing:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$$

since if $\beta_1 = 0$, the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis Testing - Continued

- To test the null hypothesis, we compute a t statistic, given by:

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

- This statistic has a t distribution with n – 2 degrees of freedom, assuming that $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing a value equal to or greater than |t|. This probability is called the p-value.

Results on Cigarette Data

	Coefficient	Std. Error	t Statistic	p-value
Intercept	148.624	2.660	55.88	$< 2 \times 10^{-16}$
taxs	-0.913	0.055	-16.66	$< 2 \times 10^{-16}$

Assessing the Overall Accuracy of the Model

- We calculate the Residual Standard Error (RSE):

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

where the residual sum of squares is:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 .$$

Results on Cigarette Data

Quantity	Value
Residual Standard Error	18.73
R^2	0.3453
Adjusted R^2	0.344
F Statistic	277.4

Multiple Linear Regression

- The model is:

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

- 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 cigarette consumption example, the model becomes:

$$\text{packpc} = \beta_0 + \beta_1 \times \text{taxs} + \beta_2 \times \text{income\_pc} + \epsilon$$

Interpretation of Regression Coefficients

- The ideal scenario is when the predictors are uncorrelated:
  - Each coefficient can be estimated and tested separately.
  - Interpretations like "a one-unit change in X_j is associated with a β_j change in Y, while all other variables remain fixed" are possible.

Estimation and Prediction for Multiple Regression

- Given the estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula:

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

- We estimate $\beta_0, \beta_1, \ldots, \beta_p$ as the values that minimize the sum of squared residuals:

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

- This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple linear regression coefficient estimates obtained using the least squares method.

Results for Cigarette Data - Multiple Regression

	Coefficient	Std. Error	t Statistic	p-value
Intercept	133.375	5.355	24.91	$< 2 \times 10^{-16}$
taxs	-0.990	0.059	-16.72	$< 2 \times 10^{-16}$
income_pc	1.349	0.412	3.27	0.00114

Summary Statistics

Statistic	Value
Residual Standard Error	18.56 on 525 DF
Multiple R^2	0.3584
Adjusted R^2	0.3559
F-statistic	146.6 on 2 and 525 DF
p-value	$< 2.2 \times 10^{-16}$

Some Important Questions

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful for predicting the response?
2. Do all the predictors help to explain Y, or is only a subset useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is at least one predictor useful?

- To answer the first question, we can use the F statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)} \sim F_{p,n-p-1}$$

Quantity	Value
Residual Standard Error	18.56
R^2	0.3584
Adjusted R^2	0.3559
F Statistic	146.6

Assessing the Overall Accuracy of the Model: Comparison between R^2 and R^2 adj

- The R^2 is calculated as:

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

where:

  - $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the residual sum of squares.
  - $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.

- The R^2 *adj* takes into account the number of predictors and the number of observations, and is calculated as:

$$R^2_{adj} = 1 - \frac{RSS/(n - p - 1)}{TSS/(n - 1)}$$

where:

  - n is the number of observations.
  - p is the number of predictors in the model.