Dipartimento di *Scienze Politiche*
Cattedra di *Statistica*

LA CORRUZIONE IN ITALIA: UN'ANALISI QUANTITATIVA

RELATORE
Prof. Roberto Rocci

CANDIDATA
Valeria Rossi
Matricola:072202

In collaborazione con Confindustria-CeFOP LUISS

ANNO ACCADEMICO

2014 / 2015

ABSTRACT

The aim of this paper is to explain, as much as possible, with statistical models, the phenomenon of corruption in Italy. To measure and understand the level of corruption in a country is extremely relevant, considering its effect on economic and development fields. Only when the causes of corruption have come out, corruption can be defeated. That's why the fight against corruption has to start by identifying the causes of corruption.
The paper is divided into three parts. The first part is theoretical; (here) the statistical model that is going to be implemented is explained. The second part is the description of theories on the causes of corruption. The third part is the application of the model to the hypothesis given by literature.

CHAPTER 1: THE SIMPLE LINEAR REGRESSION

The simple linear regression is used to analyse the linear dependence of two variables. This gives the possibility to analyse the empirical relationship between quantitative variables that could represent social phenomena.
The simplest relationship of dependency is a linear one. A linear function is given by the formula: *f (x): y* = α + β*x*. The function predicts a value of *y* for a given value of *x*. The closer that prediction is to the reality, the more that model is significant.
To evaluate the hypotheses that there is a linear relation between two variables the first step is to represent *n* points, combination of *x* and *y*, on the Cartesian coordinate system, forming a scatterplot. A first analysis to understand if the hypothesis of linear relation is verified is visual: if the points are around an imaginary line there might be a linear relation. The second step is to estimate the imaginary line $y = \alpha + \beta x + \varepsilon$, where ε is an error term such that $E(y) = \alpha + \beta x$. The line is estimated by the prediction equation that is the best straight line in the sense that is the one that falls closest to the points in the scatterplot. This is guaranteed by the ordinary least squares method (OLS) that estimates the regression coefficients with a formula that makes the sum of least squares minimum. The sum of least squares is the sum of the difference between the y_i of the data and the $\hat{y}_i$ of the line squared: $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. The intercept and the coefficient of the prediction equation are calculated on the aim of minimizing *SSE*:

$$b = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2},$$

$$a = \overline{y} - b\overline{x}.$$

Having all the elements it can formulated the estimated expected value of y as:
$E(y) = a + b$x, this function is called linear regression function.

The way to measure the strength of the association between variables is by calculating the correlation coefficient. This is like a standardized regression coefficient since it is computed as $r = \left(\frac{s_x}{s_y}\right) b$, where s_x and s_y are the standard deviations of x and y and b the slope. Perfect correlation is when $r = \pm 1$ and none correlation is when $r = 0$, in fact when r is 1 it means that the points of the scatterplot are exactly points part of the line and $r = b$ while when $b = 0$ then $r = 0$, and this means that there is not a linear increasing or decreasing trend in the relationship. The ratio of the explained sum of squares to the total sum of squares, can be obtained by computing the determination coefficient also known as R-squared, that turn out to be equal to the square of the correlation coefficient. It indicates which part of the variation of y comes from the linear relation with x. To verify if a linear dependence between y and x exists, it is necessary to set up a test of hypotheses. The two hypotheses to test are the null and the alternative, the null hypothesis states the linear independence between the two variables and the alternative one states the opposite. The test statistic t is equal to $t = \frac{b-\beta}{se}$, where se is the estimate of the standard error of b and β is zero, considering the fact that linear independence is when the regression coefficient is zero. t is a measure of how far is b from 0 (linear independence) in terms of se. The test statistic has a sample distribution that is a Student t with a specific number of degrees of freedom (df). To reject the null hypotheses the p-value (i.e. the probability, under the null, of obtaining a value of the test statistic more extreme than the one observed) has to be small enough (under 0.05, taking a level of confidence of 0.95).

If a single variable is not enough to explain a phenomenon y, there is the possibility of considering two or more explicative variables, trough a multiple linear regression model. An example of a model with two variables is:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Each variable has a different impact on the y and this is given by their own regression coefficient. Also in this case the OLS method is employed to estimate the coefficients.

Also in the multiple linear regression model there are tools to measure the strength of the relation.

To verify the actual dependence of the y to the different variables of the model a system of hypotheses needs to be set up. The null hypotheses that states the independence of the different x towards y is given if all their regression coefficient are equal to 0, the alternative hypotheses is that at least one is different from 0.
The test statistic is:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

The F distribution has a different shape for every *df*. The bigger the R^2 is, the bigger F will be. Statistic evidence to reject the null hypotheses (of independence) grows with the increasing of F.

In the multiple linear regression model there could also be relations between the independent variables their selves. So when a variable results not significant in a model, the reason isn't necessarily that it doesn't influence y, but could be that its influence is already explained by another variable in the model.

CHAPTER 2: THE ANALYSIS OF THE CAUSES OF CORRUPTION

Corruption is a phenomenon that has a great negative impact on the economy especially on the development of a country; this is proved by some statistics on the relation between the level of corruption of a country and the technology capability of a country. The relation is negative, in the sense that the correlations are negative. So these two phenomena cannot go together.
Corruption in Italy is becoming systemic. It has deep influences on the society and on the economy as a whole. In the last decades corruption phenomenon has been increasing more and more. A quantitative description of corruption is relevant to both criminologists and jurists, in fact it is the first step to analyse the phenomenon in qualitative terms. The empirical corruption is the corruption crime. To analyse the causes of corruption crime the different theories that come from literature will be considered. Having this basis it is possible to proceed with the empirical analysis.
The definition of corruption is itself unclear, since there are many definitions and different conceptions of it. The most notable definition is the one given by the *World Bank* (1997) that considers corruption as "the abuse of public office for private gain". In many other definitions there is a stress on the role

of the State. The public actor can be anyone, from a political leader to a civil servant. The private actors are different; they can be either individuals or corporations. Even if also within the privates there can be corruption, is the one between the public and the private that has the biggest relevance and impact. The analysis of this paper will be based on the definition given by Susan Rose-Ackerman (1975) that considers the corruption as the illicit passage from public to private sphere.

To measure the corruption index the analysis will be focused on the number of corruption crimes. However, there are also other ways to measure it, such as the one used by Transparency International (1999), which considers the corruption perceived by the people as a measure of the level of corruption of a country.

There are many theories that explain the causes of corruption. Considering the fact that corruption is a crime, its commission depends on the basis of a benefit-cost consideration. The cost of committing an act of corruption is the judicial punishment and the political retaliation. In fact in a democratic system the charges of corruption is a mean to attack a political opponent. Lipset (1960) considers the political participation as a mean to control the officials' political behaviour. The political participation grows with a high level of education and with a high economic standard. Another set of theories considers that the benefit from corruption for political officials comes from the private gain taken illegally through a public service. So the more the public sector is involved the more opportunities to implement corruptive conducts come out.

Moreover, there is a contribution on the research of the causes of corruption that comes from Del Monte and Papagni (2007). Their estimations are based on economic variables and cultural variables, such as the govern expenditure and the economic development or the political fragmentation and social activities of volunteers. Unfortunately, this analysis has an incomplete data set for some regions. For that reason this paper considers the hypotheses and tries to verify them with another data set.

A further hypothesis comes from my personal considerations. The inefficiency of the entrepreneurial activities can have a great impact on the level of corruption of a territory. This comes from the consideration that inefficiency produces a lack of transparency on those activities, attracting corruptors.

The first step for a quantitative analysis of the causes of corruption, based on these theories, is to find a summary measure of corruption. Transparency International (1999) has determined a corruption perception index, but this has some inconvenient, since the definition of corruption is itself unclear. Another way to measure corruption is considering the judiciary data of crimes.

CHAPTER 3: QUANTITATIVE ANALYSIS OF CORRUPTION PHENOMENON

The quantitative analysis is based on a provincial data set taken from a Report of attractiveness and competitiveness of Italian territories (2014).

In our study the corruption is measured by the mean of the standardized (also in terms of population) judiciary variables:

1) Number of convicted for usury crime
2) Number of convicted for conspiracy
3) Number of convicted for illicit public procurement
4) Number of convicted for illicit authorizations and concessions

Considering a classification of the Italian provinces the one with the highest level of corruption is *Ragusa* and the one with the lowest level of corruption is *Bolzano*.

Considering the independent variables related to the literature given on the second chapter there is a set of variables for each hypothesis. For example for the hypotheses of Lipset (1960) there are variables that express economic attributes, such as GDP per capita and variables that express education level in the provinces such as the number of students or the expense of the province for instruction. Going trough, other variables can be identified for each of the other hypothesis: for the one that relates corruption with the intervention of the State in the economy there are variables that consider the contribution of the State to business. For the hypothesis of Del Monte and Papagni (2007) the economic development can be measured by the GDP variation from one year to another. Considering my idea of the influence of efficiency on the level of corruption, data that mirrors inefficiency on the entrepreneurial net, such as the timing (in days needed) and the cost of starting a business or the difficulty of finding qualified staff can be analysed.

Using the statistical software, GRETL, it is possible to implement a linear regression model and to estimate it by OLS (Ordinary Least Square) to identify which variables are significant and which are not on the corruption index, derived from the judiciary data. Not only does the analysis find out the significance of each variable, but also their associated regression coefficients. The method used to eliminate step by step those variables that result not significant is called “try backward” and consists on taking away the less significant variable and setting up a new OLS model with the rest of the variables.

After 19 steps, the model obtained consists of 4 variables: GDP per capita, initial funds for new enterprises, public expenditure on education and difficulty on finding qualified staff for new enterprises.
The variable GDP per capita has a negative coefficient, the variable initial fund for new enterprises has a negative coefficient, the variable public expenditure on education has a positive coefficient and finally the variable difficulty on finding qualifies staff for new enterprises has a positive coefficient.
The first analysis can be given from the coefficients (b) of each variable in the multiple linear model of regression settled up. Considering them, the variables are not all coherent with the literature. In fact the expense on education has a negative coefficient instead of the expected positive one as the variable "initial funds". As mentioned earlier, within a multiple model the possible effect of an independent variable on y can be hidden from other variables that already explain this effect. So a further analysis is to consider a simple linear regression model for each of the variables that are part of the multiple linear regression model. With this analysis it can be observed how the variables alone affect the variable y. Actually, considering the variable "funds for new enterprises" it has a positive coefficient. This is coherent with the literature. Instead regarding the variable "expense on education", the relation is still opposite to the one considered by the literature. The positive relation means that the growth of such expense makes the index of corruption higher. This goes against the thesis that the growth of the level of education expressed by the expense on it leads to a decrease of the index of corruption. This doesn't induce the conclusion that the thesis is wrong. Indeed, the result probably comes out from different reasons. For example one of the reasons can be identified by the fact that the expense on education is not a real measure of the level of education of a province, but it is a proxy of the general expenses of each province. The other variables, "difficulty on finding qualified staff for new enterprises" and "GDP per capita" are still coherent by their selves with the literature on them. On the one hand the variable "difficulty on finding qualified staff for new enterprises" has a positive correlation: the harder it is to find qualified staff, the less efficient are the enterprises' system; So, as a consequence, the enterprises' environment more unclear and cloudier attracting in this way corruption. On the other hand "GDP per capita" has a negative impact on the index of corruption in fact, as Lipset (1960) considered, the more high is the economic level of the population, the more active their political behaviour is, specially during elections, controlling the operate of public officials efficiently, avoiding more easily corruption practices.

To sum up, the statistical model settled up has verified some of the hypothesis taken from literature. A coefficient of determination (R-squared) of 0.53 for such a complex social phenomenon as corruption is high enough to consider relevant the hypothesis examined.

SOMMARIO

INTRODUZIONE

Scopo della seguente tesi è spiegare, per quanto possibile, il fenomeno della corruzione in Italia con l'ausilio di modelli statistici. Il livello di corruzione in un Paese è decisivo, per gli effetti significativi su sviluppo ed economia. La corruzione può essere vista come una tassa occulta che impoverisce il paese. La lotta alla corruzione deve essere intrapresa partendo dall'identificazione delle sue cause. Solo agendo su queste, vi può essere la possibilità di agire contro tale fenomeno. Secondo le parole di Raffaele Cantone, presidente dell'ANAC (Autorità Nazionale Anticorruzione) "Il momento in cui si compie il reato di corruzione è solo l'ultimo passaggio di un effetto cascata che parte dal complesso sistema del malaffare italiano: clientelismo politico, cricche, mancanza di trasparenza e di controlli". In effetti la difficoltà di combattere il reato di corruzione sta proprio nella complessità del tessuto nel quale è immerso. Per poter analizzare il fenomeno della corruzione, e soprattutto le sue cause, bisogna dunque considerare diverse dimensioni, che vanno dal livello culturale a quello economico. L'obiettivo di questa tesi è proprio quello di addentrarsi nelle diverse dimensioni sociali per verificare la consistenza di ipotesi tratte da diverse note teorie sulle cause del fenomeno di corruzione.
La tesi è divisa in tre capitoli. Il primo capitolo è di tipo teorico: in esso è illustrato il modello statistico adottato per l'analisi quantitativa del fenomeno della corruzione. Il secondo capitolo riporta le teorie più note e le considerazioni personali sulle cause della corruzione, infine, il terzo capitolo applica alle teorie il metodo statistico illustrato nel primo capitolo.

CAPITOLO 1
REGRESSIONE LINEARE SEMPLICE

La regressione lineare semplice serve per analizzare la dipendenza di una variabile *y* da un'altra *x.* L'obiettivo dell'utilizzo di tale tecnica di analisi è investigare le relazioni empiriche tra variabili per poter analizzare le cause x_i che possono spiegare un determinato fenomeno *y*.

1.1 RELAZIONE LINEARE

Il modo più semplice con cui una variabile *y* dipende da un'altra *x* è LINEARMENTE. La funzione lineare è data dalla formula: *f(x): y* = α + β*x*. Dove α è l'intercetta e β è il coefficiente angolare. Per interpretare l'intercetta si considera il caso in cui *x* è 0, f(0): y = α, questo vuol dire che la retta interseca l'asse delle *y* (intercetta *y*) nel punto (0, α) del piano cartesiano. Il coefficiente angolare β è uguale al cambio, in unità, di *y*, dato un incremento unitario di *x*. Quando *x* è 0 allora *y* è α, quando *x* è 1 y = α + β. Questo vuol dire che quando *x* cresce di un'unità, *y* cresce di β unità. Per disegnare la retta che rappresenta la relazione tra le due variabili *y* e *x,* bastano due punti appartenenti a tale retta.

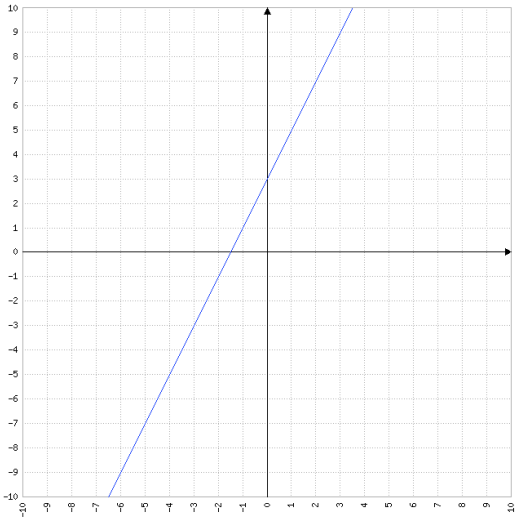
Esempio:

y = 3 + 2x:

a) 3 è l'intercetta y
Quando x=0 allora y=3

b) *2 è il coefficiente angolare della retta* (x = 0, y = 3; x = 1, y=5 → incremento di y dato da un incremento unitario di x: y = 5 – 3 = 2)

c) *A: (0,3) B: (1,5)*
Una sola retta passa per A e per B ed è la seguente:



Un modello è una semplice approssimazione della relazione tra variabili in una determinata popolazione. La funzione lineare è la funzione matematica più semplice. Il modello $y = \alpha + \beta x$ predice un valore della *y* per una data *x*. Più tale predizione si avvicina alla realtà, più il modello è significativo.

1.2 METODO DEI MINIMI QUADRATI APPLICATO ALLA RETTA TEORICA

Su un campione di *n* unità statistiche sono osservati i valori relativi a due variabili *x* e *y*. Ad esempio su *n* individui si rilevano i dati relativi a *x*: anni di educazione e *y*: reddito annuale. L'ipotesi iniziale è l'esistenza di una relazione lineare tra le due variabili *x* e *y* ovvero ci si chiede se gli anni di educazione di un individuo siano direttamente proporzionali al suo reddito annuale. In primo luogo si esegue un'analisi grafica. I dati disponibili vengono rappresentati graficamente su un piano cartesiano, date le *n* osservazioni si avranno *n* punti (*x*, *y*) che formeranno un *grafico a dispersione (scatterplot)*. Se dallo *scatterplot* si può cogliere una tendenza dei punti ad approssimarsi a una retta, si prosegue con l'analisi. Altrimenti si rigetta l'ipotesi di relazione lineare tra le due variabili (*x* e *y*). Quando dallo *scatterplot* il modello lineare $y = \alpha + \beta x$ sembra essere realistico, si possono utilizzare i dati campionari disponibili per stimare la retta; dalla funzione lineare stimata discendono *valori teorici* di *y* per ogni *x* scelta. Per garantire la massima vicinanza della retta teorica ai dati osservati, si utilizza il metodo dei minimi quadrati ordinari per stimare i due coefficienti di regressione (α e β). Il metodo dei minimi quadrati consiste nel misurare la vicinanza della retta teorica ai dati osservati attraverso la differenza tra il valore osservato e quello stimato in termini assoluti, detta errore campionario o residuo campionario. Nello scatterplot il residuo campionario di una data osservazione è la distanza

verticale del punto dalla retta teorica. Ogni osservazione ha un residuo campionario. Più la retta teorica è vicina ai vari punti, più i residui sono piccoli. La grandezza dei residui può essere sintetizzata dalla somma dei loro quadrati. La somma dei quadrati dei residui (*SSE*) è data dalla seguente formula:

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2. \qquad 1.2.1$$

Il residuo campionario è quindi calcolato per ogni osservazione del campione, ogni residuo è elevato al quadrato, infine SSE è dato dalla somma di tutti i quadrati.
Per minimizzare la *SSE* si stimano il coefficiente angolare e l'intercetta di *y* con le seguenti formule:

$$b = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}, \qquad 1.2.2$$

$$a = \overline{y} - b\overline{x}, \qquad 1.2.3$$

$\overline{x}$: è la media della variabile *x,*
$\overline{y}$*:* è la media della variabile *y.*

La retta teorica $\hat{y} = a + b\hat{x}$, con i valori stimati di a e b, ha il valore più piccolo possibile di SSE rispetto a tutte le possibili rette teoriche, grazie al metodo dei Minimi Quadrati Ordinari – OLS (acronimo dall'inglese Ordinary Least Squares) che minimizza *SSE* attraverso la stima di *a* e *b*.

1.3 IL MODELLO DI REGRESSIONE LINEARE

Prendendo in considerazione il modello $y = \alpha + \beta x$, ad ogni *x* corrisponde una ed una sola *y*. Questo genere di modello è definito *deterministico.* In una ricerca nell'ambito delle scienze sociali tale modello non è realistico in quanto, soggetti caratterizzati da un determinato valore di *x*, non avranno sempre lo stesso valore di *y*. Riprendendo l'esempio precedente, dove *x* = numero di anni di educazione e *y* = reddito annuale, tutti i soggetti che hanno ricevuto 12 anni di educazione (*x* = 12) non avranno lo stesso reddito annuale. Tuttavia utilizzando una distribuzione di probabilità condizionata a valori fissati di *x* è possibile dimostrare come una funzione lineare è la base

di tale modello. Una funzione di probabilità che descrive il reddito annuale per tutti i soggetti che hanno ricevuto 12 anni di educazione è una distribuzione condizionata. Le diverse distribuzioni condizionate originate da diversi valori di *x* avranno medie rispettivamente diverse. La funzione che descrive l'andamento di tali valori medi, dati i diversi valori di *x*, sarà una funzione lineare.
Una funzione di probabilità utilizza $\alpha + \beta x$ per rappresentare i valori medi di *y* in funzione di *x.*

$E(y) = \alpha + \beta x$.

Per un valore di x dato, α + βx rappresenta la media della distribuzione condizionata di y per i soggetti aventi tale x.

E (y) : *Valore atteso di y (Expected value).*

Una funzione che mette in relazione i valori della *x* e la media della distribuzione condizionata di *y* è chiamata *Funzione di Regressione.* Tale funzione in questo caso è lineare e graficamente è espressa da una retta; β è il coefficiente di regressione della funzione di regressione lineare. I parametri della funzione di regressione lineare non sono conosciuti, ma vengono stimati dal metodo dei minimi quadrati, precedentemente illustrato. Per un valore di *x*, $\hat{y} = a + bx$ stima il valore atteso di *y* per tutti i soggetti della popolazione che hanno tale valore di *x.*

Il modello di regressione lineare ha un ulteriore parametro σ che descrive la deviazione standard di ogni distribuzione condizionata. σ misura la variabilità dei valori di *y.dato* un certo *x*.
Generalmente si assume che la distribuzione di *y* è *normale* se condizionata ad un valore di *x*. Il modello di regressione lineare assume che la deviazione standard della distribuzione condizionata di *y* non dipende *x*.
La stima di σ è:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-2}}. \qquad 1.3.1$$

Questo stimatore presenta una correzione per i gradi di libertà, giacché a denominatore è riportato il numero delle osservazioni meno il numero dei regressori.

1.4 LA CORRELAZIONE

Il modello di regressione lineare utilizza una retta per descrivere una relazione. Informazioni sull'associazione tra due variabili sono fornite da due misure.

1. Il coefficiente angolare (o coefficiente di regressione)
2. L'indice di correlazione

1. Il coefficiente di regressione b può fornire informazioni riguardanti la direzione dell'associazione di due variabili x e y: al crescere di x, y cresce, ovvero, al crescere di x, y decresce. Tuttavia il coefficiente di regressione non fornisce informazioni riguardanti la forza dell'associazione tra due variabili x e y. Questo in quanto il suo valore numerico è intrinsecamente connesso all'unità di misura, dunque può essere piccolo o grande a piacere a seconda dell'unità di misura scelta.

2. L'indice di correlazione è una versione *standardizzata* del coefficiente angolare. Il suo valore non dipende da alcuna unità di misura. L'indice di correlazione è il valore che il coefficiente angolare avrebbe se le variabili x e y avessero le stesse deviazioni standard. Le deviazioni standard di x e y sono rispettivamente s_x e s_y

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}, \qquad 1.4.1$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}. \qquad 1.4.2$$

La relazione che intercorre tra l'indice di correlazione r ed il coefficiente angolare b della retta teorica $\hat{y} = a + bx$ è data da:

$$r = \left(\frac{s_x}{s_y}\right) b. \qquad 1.4.3$$

Quando s_x e s_y sono uguali, $r = b$. Ad esempio, quando le variabili x e y sono standardizzate, entrambe avranno deviazioni standard uguali a 1. Data la relazione tra r e b, l'indice di correlazione è anche chiamato "coefficiente di regressione standardizzato" del modello $E(y) = \alpha + \beta x$. Nella pratica non è necessario standardizzare le variabili, ma è utile interpretare l'indice di

correlazione come il valore del coefficiente angolare quando le deviazioni standard delle variabili sono uguali.
Le proprietà dell'indice di correlazione sono le seguenti:

- È una misura della forza dell'associazione valida solo quando la retta è un modello adatto per la relazione tra le variabili *x* e *y*. Dato che r è proporzionale al coefficiente angolare della retta teorica, esso misura la forza dell'associazione lineare tra *x* e *y*.

- $-1 \leq \mathrm{r} \leq 1$.

- r ha lo stesso segno del coefficiente angolare. Dato che r è uguale a b moltiplicato per il rapporto delle due deviazioni standard (sempre positive), il segno rimane invariato. Dunque, quando $b > 0$ le variabili sono correlate positivamente, quando $b < 0$ sono correlate negativamente.

- $r = 0$ per le rette che hanno il coefficiente angolare pari a 0. In tal caso non vi è né incremento né decremento lineare nella relazione tra *x* e *y*.

- $r = \pm 1$ quando i punti dello *scatterplot* coincidono perfettamente con la retta teorica. In tal caso non vi è alcun errore campionario (residuo campionario).

- Più è grande il valore assoluto di *r*, più è forte l'associazione lineare tra *x* e *y*.

- Il valore di *r* non dipende dall'unità di misura delle variabili.

1.5 IL COEFFICIENTE DI DETERMINAZIONE

Il coefficiente di determinazione indica di quanto la varianza totale di *y* derivi dalla dipendenza lineare tra *y* e *x.* Tale informazione indica l'aderenza del modello al fenomeno di studio. Per questo motivo il coefficiente di determinazione è chiamato anche indice di bontà di adattamento. *TSS*, la devianza totale di *y,* è composta dalla devianza spiegata dalla regressione *ESS* e la devianza residua *RSS*. In particolare:

$$TSS = ESS + RSS \qquad 1.5.1$$

$$TSS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \qquad 1.5.2$$

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad 1.5.3$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad 1.5.4$$

Per determinare quanta parte della devianza totale sia spiegata dalla retta di regressione si può considerare il coefficiente di determinazione così calcolato:

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} \qquad 1.5.5$$

Le proprietà del coefficiente di determinazione sono le seguenti:

- Quando r^2 è uguale a 1 la devianza residua è nulla. I punti giacciono sulla retta di regressione, dunque il modello lineare stimato spiega completamente il fenomeno studiato.

- Quando r^2 è uguale a 0 la devianza di regressione è nulla e indica la mancanza di relazione lineare tra *x* e *y.*

- r^2 non fornisce informazioni sulla direzione della relazione tra *y* e *x*, bensì fornisce informazioni sulla dispersione dei punti intorno alla retta di regressione.

- A seconda del contesto si stabilisce quanto grande è il valore di r^2 per poter considerare la retta di regressione un buon modello della distribuzione *y*.

- $\boldsymbol{r^2 = (r)^2}$ il coefficiente di determinazione è uguale al quadrato dell'indice di correlazione.

1.6 INFERENZA PER IL MODELLO LINEARE

L'inferenza statistica per la regressione si basa su determinate assunzioni:

- L'analisi utilizza la randomizzazione. La y_i, per un dato valore di *x*, segue una distribuzione di tipo normale.

- La media di *y* è legata al valore di *x* dalla seguente funzione lineare E(y) = α + βx.

- La deviazione standard di *y* condizionata a *x,* indicata con σ, è identica per ogni valore di *x.*

Nell'ipotesi che la media della popolazione di *y* sia identica per ogni valore di *x,* in altre parole, che la distribuzione condizionata normale di *y* sia uguale per ogni valore di *x.* In tal caso, le due variabili quantitative *x* e *y* sono indipendenti. La funzione di regressione lineare
E(y) = α + β*x* avrà coefficiente di regressione β pari a 0. Per verificare l'ipotesi di indipendenza lineare delle due variabili si imposta un sistema di ipotesi dove H_0 è l'ipotesi nulla, mentre H_1 l'ipotesi alternativa:

H_0: β = 0 , *x* e *y* sono linearmente indipendenti
H_a: β ≠ 0 , *x* e *y* non sono linearmente indipendenti

La statistica test è uguale a

$$t = \frac{b}{se} \ , \qquad 1.6.1$$

dove *se* è l'errore standard del coefficiente angolare campionario *b* stimato come

$$se = \frac{s}{\sqrt{\sum(x_i - \overline{x})^2}} = \frac{s}{(n-1)s_x^2} \ . \qquad 1.6.2$$

Si Utilizza lo stimatore *b* per il parametro β meno l'ipotesi nulla (β = 0) e si divide per l'errore standard dello stimatore *b.* Date le precedenti assunzioni, questa statistica test ha distribuzione campionaria *t* con *df* (gradi di libertà) = *n*-2. I gradi di libertà sono gli stessi della deviazione standard condizionata dello stimatore di *s*.
Precisamente *b* stima β. Si ha un valore piccolo di *s* quando i punti dello *scatterplot* si discostano di poco dalla retta teorica di regressione. Inoltre l'errore standard di *b* è inversamente proporzionale a s_x^2 , la varianza di *x*; Quindi più la variabilità di *x* cresce, più *b* stima precisamente β.

Per rifiutare l'ipotesi nulla, cioè l'ipotesi di indipendenza lineare, si ricorre all'uso del *p*-value. Il *p*-value è la probabilità di osservare valori uguali o più estremi della statistica test *t*, prendendo per vera l'ipotesi nulla. Quanto più è

piccolo il valore del p-value, tanto più aumenta il livello di confidenza con cui si rifiuta l'ipotesi nulla. Il *p*-value per l'ipotesi alternativa Ha: $\beta \neq 0$ è la probabilità a due code data dalla distribuzione *t*-student. Più è alto il numero di gradi di libertà (*df*) (dunque più *n* è grande) più la distribuzione *t* si approssima alla distribuzione *z* e il *p*-value può essere approssimato utilizzando la tavola dei valori di probabilità della normale. Un valore piccolo del *p*-value per H_0: $\beta = 0$, porta a rifiutare l'ipotesi nulla e suggerisce un coefficiente di regressione della retta di regressione diverso da 0. Bisogna ricordare che il *p*-value non determina la probabilità del verificarsi dell'ipotesi nulla. Per convenzione si considera un valore del *p*-value minore di 0.05 sufficientemente piccolo per rifiutare l'ipotesi nulla. Un intervallo di confidenza per β è dato da $b \pm t(se)$. Il valore *t* è preso dalla tavola *t*-student, con gradi di libertà pari a $df = n - 2$ ed è tale che $\Pr(T > t) = \propto/2$, dove $1-\propto$ è il livello di confidenza.

1.7 REGRESSIONE LINEARE MULTIPLA

Un modello che contiene più variabili esplicative $x_1, x_2, \ldots, x_n$ è in grado di provvedere maggiori informazioni rispetto a quelle ricavabili dal modello semplice sopra descritto, il quale è in grado di analizzare solamente due variabili alla volta.
Per semplicità suppone di avere due variabili esplicative x_1 e x_2. La funzione di regressione multipla è la seguente:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

Per determinati valori di x_1 e x_2 la funzione fornisce il valore atteso di *y* per tutti i soggetti che hanno tali valori di x_1 e x_2. Ogni variabile esplicativa ha il proprio coefficiente angolare dato che ognuna incide su *y* in maniera diversa. Per rappresentare graficamente il modello di regressione multipla non è sufficiente il piano cartesiano a due dimensioni. Infatti, nel caso di due variabili esplicative, si deve aggiungere una terza dimensione; nel caso di tre variabili, si deve aggiungere una quarta dimensione; e così via. Se si fissa un valore di x_2 la funzione è semplificata ad una retta. Il valore di x_2 è dunque controllato e il coefficiente angolare β_2 è uguale per ogni valore fissato di x_2. Nella regressione multipla un coefficiente angolare descrive l'effetto della variabile esplicativa su *y* tenendo sotto controllo gli effetti delle altre variabili esplicative del modello, mentre nella regressione semplice il coefficiente angolare descrive l'effetto dell'unica variabile esplicativa ignorando l'effetto di qualsiasi altra possibile variabile esplicativa. Per questo motivo i parametri β_1 e β_2 sono chiamati coefficienti di regressione parziali. Come nel modello di

regressione lineare semplice i parametri della funzione sono stimati attraverso il metodo dei minimi quadrati.

1.8 INDICI DI CORRELAZIONE E DI REGRESSIONE MULTIPLA

Anche nel modello di regressione multipla si utilizzano degli indici per indicare la forza di associazione tra la variabile *y* e le variabili esplicative x_1, x_2, x_3, ..., x_k. Le variabili esplicative sono considerate tutte insieme, come se agissero come un'unica variabile; la forza di associazione tra queste e *y* aumenta più i valori di $\hat{y}$ sono correlati con i valori di *y* osservati. L'indice di correlazione multipla è indicato con *R*. Ogni individuo della popolazione studiata avrà un valore di *y* osservato e un valore di $\hat{y}$ teorico generato dalla funzione di regressione multipla. I valori teorici $\hat{y}$ non possono essere correlati negativamente con i valori osservati di *y* in quanto i valori teorici devono rappresentare *y* per lo meno quanto lo può fare il valore medio del campione. Quindi, considerando che il valore medio $\overline{y}$ è uguale a $\hat{y}$ quando tutti i coefficienti angolari parziali sono uguali a 0 e che in tal caso $\hat{y}$ ha un indice di correlazione nullo rispetto a *y*, si deduce che il valore di *R* è compreso tra 0 e 1. Più è grande il valore di *R*, più il modello è rappresentativo. Un altro indice che descrive la capacità del modello a rappresentare i valori di *y* utilizza il concetto di riduzione proporzionale dell'errore. Tale indice è il corrispettivo di r^2. Questa misura indica di quanto migliore è la funzione di regressione rispetto alla media del campione nello stimare *y*. Se si stima *y* senza usare $x_1, x_2, x_3, ..., x_k$, il migliore stimatore è la media del campione, $\overline{y}$. Mentre se si usano i valori di $x_1, x_2, x_3, ..., x_k$ il miglior stimatore risulta essere $\hat{y}$, ricavato dalla funzione di regressione. Gli errori di stima vengono calcolati dalla differenza dei valori osservati e quelli stimati di *y*. Nel caso in cui si usi lo stimatore $\overline{y}$, l'errore di stima è y-$\overline{y}$ mentre se si utilizza lo stimatore $\hat{y}$ l'errore è y-$\hat{y}$. In entrambi i casi si sintetizza l'errore attraverso la somma dei quadrati degli errori di stima come nelle equazioni 1.4.2 e 1.4.3. R^2 misura la proporzione della variazione totale in *y* spiegata dal potere di stima di tutte le variabili esplicative attraverso il modello di regressione multipla. Il simbolo suggerisce che corrisponde al quadrato dell'indice di correlazione multipla *R*. Si calcola nel modo seguente:

$$R^2 = \frac{TSS - SSE}{TSS}. \qquad 1.8.3$$

Le formule di R^2 e di r^2 sono identiche, r^2 è in realtà un caso specifico di R^2 applicato al modello di regressione con un'unica variabile esplicativa.

Il modello di regressione multipla deve essere uno stimatore migliore non solo dello stimatore $\overline{y}$, ma anche dei singoli modelli di regressione semplice generati da ogni variabile esplicativa.
Le proprietà di R^2 sono le seguenti:

- $0 \leq R^2 \leq 1$.

- Più è grande il valore di R^2, più le variabili esplicative $x_1, x_2, x_3, \ldots, x_k$ collettivamente stimano *y.*

- $R^2=1$ quando la funzione di regressione multipla passa attraverso tutti i punti del campione, in quanto $SSE=0$ ovvero $y=\hat{y}$.

- $R^2=0$ quando le stime non variano al variare delle *x*. In tal caso i coefficienti angolari stimati sono tutti uguali a 0 e $\hat{y} = \overline{y}$.

- R^2 non può decrescere quando si aggiungono ulteriori variabili esplicative, in quanto è impossibile spiegare minore variazione in *y* aggiungendo variabili esplicative in un modello di regressione.

- R^2 per la regressione multipla è per lo meno grande quanto gli indici r^2 per i modelli di regressione semplice delle singole variabili esplicative.

Non si possono ignorare le relazioni tra le variabili esplicative. Il modello ideale sarebbe composto da variabili esplicative non correlate tra loro ma fortemente correlate con *y.* Tale modello, in pratica, non è sempre possibile. Talvolta, all'aggiungere una nuova variabile esplicativa (ad esempio x_3), il valore R^2 cresce di poco: questo non necessariamente deriva dall'assenza di correlazione tra la nuova variabile x_3 e la *y*, ma può derivare da una forte correlazione tra la nuova variabile e una delle variabili esplicative già presenti nel modello; il contributo di x_3 è già dato dalla variabile a cui è fortemente correlato.

1.9 INFERENZA PER IL MODELLO LINEARE MULTIPLO

L'inferenza statistica per il modello lineare multiplo si basa su determinate assunzioni:

- Il campione di *y* è scelto attraverso la randomizzazione.

• La distribuzione della popolazione di y è di tipo normale per ogni combinazione di valori di $x_1, \dots, x_k$.

• La deviazione standard condizionata, σ, è uguale per ogni combinazione di valori di $x_1, \dots, x_k$.

Per verificare l'effetto delle variabili $x_1, \dots, x_k$ prese collettivamente sulla variabile risposta si imposta il seguente sistema di ipotesi:

H_o: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$, ovvero, il valore atteso di y non dipende dai valori di $x_1, \dots, x_k$.

H_1: Almeno un $\beta_i \neq 0$, ovvero, il valore atteso di y dipende da almeno uno dei valori di $x_1, \dots, x_k$.

Tale sistema di ipotesi può essere definito anche come:

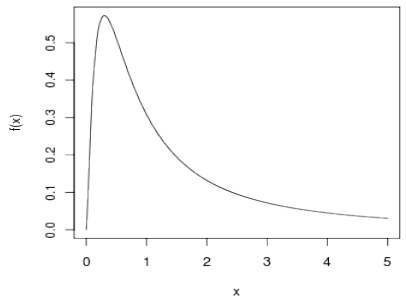
H_0: $Correlazione\ multipla\ nella\ popolazione = 0$,

H_1: $Correlazione\ multipla\ nella\ popolazione > 0$.

Seguendo tale sistema di ipotesi, la statistica test è la seguente:

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}. \qquad 1.9.1$$

La distribuzione di tale statistica è chiamata *distribuzione F* .



La *distribuzione F* può assumere solo valori positivi ed è schiacciata sul lato destro.
La forma della *distribuzione F* è determinata da due termini di gradi di libertà, df_1 e df_2:
$df_1 = k$, il numero di variabili esplicative nel modello.
$df_2 = n - (k + 1) = n$ – *il numero dei parametri nella funzione di regressione.*
df_1 è il divisore del numeratore di F; df_2 è il divisore del denominatore di F.
La media della distribuzione si approssima a 1.

Più è grande il valore di R^2 , più la statistica F è grande. Al crescere del valore di F, cresce l'evidenza statistica che porta a rifiutare l'ipotesi nulla H_0. Prendendo per vera l'ipotesi nulla, il p-value è la probabilità che la statistica test F sia più grande del valore osservato di F. Questa probabilità corrisponde all'area sottostante alla coda destra della *distribuzione F* dal valore del test F. È dunque l'area sottostante alla funzione nella zona maggiore del valore del test F.

Come detto in precedenza, al crescere del numero di variabili, il modello di regressione multipla diventa più difficile da interpretare e alcune variabili esplicative potrebbero risultare ridondanti, specialmente quando sono fortemente correlate con altre variabili esplicative del modello. È possibile verificare se un *modello completo* è più efficace di una sua semplificazione (*modello ridotto*) o meno. Il modello ridotto è un caso particolare del modello completo. Il modello completo e quello ridotto sono identici se i coefficienti di regressione parziale per le variabili extra, nel modello di regressione completo, sono uguali a 0. Verificare se il modello completo è uguale al modello ridotto può esser fatto attraverso un sistema di ipotesi. Se il modello completo e il modello ridotto sono i seguenti:
-modello completo con tre variabili esplicative e tutti i termini di interazione di secondo ordine:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3,$$

-modello ridotto senza termini di interazione:

$$E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

il sistema di ipotesi è il seguente

H_0: $\beta_4 = \beta_5 = \beta_6 = 0$

H_1: almeno uno tra β_4 , β_5 e β_6 è $\neq 0$

Basta anche solo uno dei tre coefficienti ad essere diverso da 0 per rifiutare H_o.
La statistica test che paragona due modelli di regressione, paragona la somma dei quadrati residui *SSE*. Si chiami la *SSE* del modello completo SSE_C e del modello ridotto SSE_r.
$SSE_c > SSE_r$ in quanto il modello ridotto ha meno termini e tende di base a produrre stime meno buone, come detto nel paragrafo 1.8.
La statistica test è uguale a:

$$F = \frac{(SSE_r - SSE_c)/df_1}{SSE_C/df_2} = \frac{(R_c^2 - R_r^2)/df_1}{(1-R_c^2)/df_2}, \qquad 1.9.2$$

df_1: numero di termini in più del modello completo rispetto a quello ridotto, nel caso in particolare 3.

df_2: *n-(k+1)* = *n* – il numero dei parametri nel modello completo.

Una grande diminuzione dell'errore (o un grande incremento di R^2) genera un valore grande del test *F* e un valore piccolo del *p*-value. Come già visto nel paragrafo 1.8, il *p*-value sarà la probabilità data dalla coda destra della *distribuzione F*.

CAPITOLO 2

ANALISI DELLE CAUSE DELLA CORRUZIONE

2.1 L'IMPORTANZA DEL FENOMENO DELLA CORRUZIONE

Lo studio del fenomeno della corruzione è di fondamentale importanza. Ed è per questo che l'analisi delle sue cause è determinante se lo si vuole combattere. L'importanza della lotta alla corruzione è dimostrata dal fatto che la sua presenza ha un effetto negativo significativo sulla capacità tecnologica di un territorio. L'indice di corruzione è correlato negativamente con tutti gli indicatori di capacità tecnologica. Statisticamente parlando l'indice di correlazione è al di sopra di 0.5 per la maggior parte di questi. Come lo dimostra l'analisi "Basic versus Innovation" (2015) illustrata di seguito:

	Availability of latest techologies	**Firm-level technology absorption**	**Techological adoption**	**Entreprises having purchased online (at least 1%)**	**Entreprises having received orders online (At least 1%)**	**Enterprises with fixed broadband access**	**FDI and technology transfer**
Country level corruption perception	-0,57	-0,63	-0,54	-0,6	-0,52	-0,39	-0,17
Regional level corruption perception	-0,57	-0,63	-0,56	-0,55	-0,5	-0,38	-0,21

Secondo il rapporto del *Group of States against corruption* (GRECO) (2011) "la corruzione è profondamente radicata in diverse aree della pubblica amministrazione, nella società civile, così come nel settore privato. Il pagamento delle tangenti sembra pratica comune per ottenere licenze e permessi, contratti pubblici, finanziamenti, per esercitare la professione medica, stringere accordi nel mondo calcistico, ecc.(...) La corruzione in Italia

è un fenomeno pervasivo e sistemico che influenza la società nel suo complesso". Negli ultimi decenni il fenomeno di corruzione è divenuto di larghissima diffusione; non riguarda solo piccoli burocrati, ma anche vertici dell'amministrazione.

2.2 LA DEFINIZIONE DI CORRUZIONE

Divenendo un fenomeno sempre più rilevante, si cerca di darne una definizione quantitativa e analizzarne le cause. Una descrizione quantitativa è rivelante sia per il criminologo che per il giurista, in quanto costituisce la base indispensabile e preliminare per analizzare il fenomeno da un punto di vista qualitativo. La corruzione come grandezza empirica è il reato di corruzione. Avendo una misura del reato di corruzione si possono analizzare le sue cause, per porre le basi per l'implementazione di modelli normativi di risposta. Per analizzare empiricamente la cause della corruzione ci si baserà sulla letteratura proveniente a riguardo. Avendo delle ipotesi sulle cause della corruzione si potrà proseguire con un'analisi empirica.

Per prima cosa si deve definire il termine corruzione. La definizione più nota e più semplice è quella data dalla *World Bank* (1997) secondo la quale la corruzione è l'abuso del potere pubblico per beneficio privato. In molti casi, l'abuso di potere pubblico non necessariamente è finalizzato a benefici di un unico privato ma può essere finalizzato al beneficio di un partito, di una classe sociale o più in generale di un gruppo di persone. In generale l'abuso avviene attraverso richieste di somme di denaro (tangenti) per conseguire guadagni illeciti, ma esistono anche altri tipi di abusi che non comportano l'uso di tangenti. Ad esempio vi è un abuso di posizione quando un impiegato pubblico richiede ferie per malattia, ma in realtà non è affetto da alcuna malattia e utilizza i giorni concessi per andare in vacanza. In tale caso non vi è alcuna tangente in questione e si tratta comunque di un atto di corruzione. Il ruolo dello stato è rappresentato nella maggior parte delle definizioni della corruzione. La corruzione, secondo questo tipo di definizioni, è una particolare relazione tra stato e società. La corruzione è convenzionalmente un fenomeno in cui il potere pubblico è usato a fini privati. Nelle diverse definizioni si enfatizzano punti diversi, ad esempio, secondo Susan Rose-Ackerman (1975), tra le più note studiose del fenomeno della corruzione nel mondo, la corruzione è una transizione illegittima di attori che passano dalla sfera pubblica alla sfera privata utilizzando beni collettivi. Nell'arena nazionale la corruzione avviene nel punto di incontro tra lo stato e i vari attori non statali. Da una parte vi è l'attore statale dall'altra, il corruttore.

L'attore statale può essere chiunque, partendo dai leader politici all'impiegato pubblico. Gli attori non statali sono di diverso tipo, da un individuo a una corporazione. Esiste anche la corruzione a livello privato, tuttavia nelle definizioni di corruzione si enfatizza più la corruzione come relazione tra stato e società dato che il settore pubblico è considerato di fondamentale importanza. Effettivamente controllare il settore pubblico è un prerequisito per controllare il settore privato. Tenendo conto di questo, la definizione su cui si baserà l'analisi sarà quella descritta da Susan Rose-Ackerman(1975).

2.3 LA MISURA DELLA CORRUZIONE

Ricerche empiriche sulle determinanti della corruzione sono scarse, data la difficoltà che si riscontra nella misurazione della stessa corruzione, essendo un fenomeno nascosto e dunque spesso sommerso. Di recente economisti e scienziati politici hanno usato indici di corruzione "percepita", costruiti attraverso le risposte di cittadini e uomini di affari a questionari specifici. L'Indice di Corruzione Percepita prodotto da Transparency International (1999) colloca l'Italia nello stesso gruppo di numerosi paesi dell'Est Europa, dell'Asia e del Sud America con un punteggio di 5.2 su 10, dove 0 indica il più alto livello di corruzione e 10 il più basso. L'indice dovrebbe chiamarsi Indice di non corruzione essendo il livello più alto (10) il minimo livello di corruzione. I crimini di corruzione in Italia sono aumentati tra la metà del 1970 e il 1990, con una leggera diminuzione nel 1993 in seguito a *Mani Pulite*. La distribuzione del crimine di corruzione non è omogenea in Italia. Una volta selezionate un numero di ipotesi sulle cause della corruzione, le stesse verranno analizzate all'interno del territorio italiano.

2.4 TEORIE SULLE CAUSE DELLA CORRUZIONE

Secondo la definizione che di prenderà in considerazione, di Rose-Ackerman(1975), la corruzione è un crimine commesso dai pubblici ufficiali per guadagno personale. Definendo la corruzione come crimine, la teoria economica della corruzione segue quella economica del crimine di Gary Stanley Becker (1968) secondo la quale un potenziale criminale, in questo caso un pubblico ufficiale, mette sulla bilancia i benefici del crimine e i costi dello stesso. Maggiori sono i benefici rispetto ai costi, più il pubblico ufficiale è propenso a commettere il crimine di corruzione. Il costo più rilevante proviene dal rischio di essere scoperti e di conseguenza condannati. Tale rischio è legato sia all'efficienza del sistema legale del Paese in cui è commesso il crimine di corruzione sia al grado di protezione che gli individui

riservano ai pubblici ufficiali corrotti. Secondo tale impostazione il rischio di subire una condanna è maggiore nei sistemi democratici, dove, data la competizione politica, l'accusa di corruzione diviene un'arma nei confronti dell'avversario politico. Tale ipotesi è basata su gli studi di Seymour Martin Lipset (1960), che considera la partecipazione politica del cittadino un modo per tenere sotto controllo l'operato politico. La partecipazione politica diviene una forma di monitoraggio diretto. Più è alto il livello d'istruzione e il reddito dell'elettorato, più è attiva la partecipazione politica. Più è attiva la partecipazione, minore è il livello di corruzione.
I benefici della corruzione per i pubblici agenti vengono ricavati riallocando il guadagno illecito a beneficio di individui privati, utilizzando il potere guadagnato per scopi individuali. Tali interventi pubblici, quali tassazione e regolamenti, possono essere usati per avvantaggiare individui privati in cambio di parte del guadagno. Pertanto, più grande è il settore pubblico, maggiori sono: il numero di regolamenti economici, le opportunità di aiutare attori privati a trarre vantaggio da tali regolamenti o a raggirarli, le possibilità di adottare pratiche corrotte.
Un altro insieme di teorie sulle cause della corruzione si è concentrato sulla frammentazione etnica. Se una regione è caratterizzata da divisioni etniche, un leader politico che rappresenta un gruppo etnico tenderà ad allocare le risorse nei confronti di tale gruppo. I membri del gruppo etnico continueranno a supportare il leader del loro gruppo pur se vengono a conoscenza di sue pratiche corrotte. In più, altre forme di divisioni sociali come la disuguaglianza di reddito porta l'elettorato a concentrarsi sulla redistribuzione piuttosto che sull'onestà dei candidati. Maggiore è il livello di disuguaglianza, più elevato è il grado di corruzione.
Di recente è stata indicata da Torsten Persson e Guido Tabellini (2010) una sistematica connessione tra leggi elettorali e corruzione. Sistemi maggioritari sembrano generare un alto livello di responsabilità degli eletti, giacché è ricercato il consenso all'interno dell'elettorato piuttosto che all'interno del partito, dunque meno corruzione. Tuttavia Baron Ernst Friedrich von Liphart (1999) sostiene che il sistema proporzionale è quello a maggiore grado di responsabilità comportando politiche più vicine alle preferenze degli elettori e scoraggiando pratiche corrotte.

Come noto vi è un unico contributo empirico sull'analisi delle determinanti della corruzione nelle Regioni italiane di Alfredo Del Monte e Erasmo Papagni (2007). Le loro stime mostrano come variabili economiche e fattori politici e culturali influenzino la corruzione in Italia. Le variabili economiche considerate sono le spese del governo e il livello di sviluppo economico, quelle culturali e politiche sono date dalla frammentazione politica, la presenza di organizzazioni di volontariato e l'astensionismo nelle elezioni

nazionali. Ma tale analisi presenta un data-set regionale incompleto per tutte le variabili che sono utilizzate come determinanti della corruzione, in quanto mancano alcuni dati.
Secondo l'ampia letteratura a riguardo, la corruzione può essere determinata da diversi fattori che possono essere indicati dalle seguenti caratteristiche: livello di educazione e reddito degli individui, burocrazia, ineguaglianza sociale all'interno del Paese, leggi elettorali, spesa pubblica, livello di sviluppo economico del Paese, fattori culturali come astensionismo alle elezioni nazionali, frammentazione politica e presenza di organizzazioni volontarie. Si esamineranno in seguito tali indicatori che determinano, secondo la letteratura, un grado più alto o meno dell'indice di corruzione.

2.5 UNA MISURA SINTETICA DELLA CORRUZIONE

Una volta determinate le ipotesi sulle cause della corruzione, sulla base della letteratura disponibile, è possibile passare alla determinazione delle variabili quantitative da considerare.
Per prima cosa si deve identificare una misura sintetica della corruzione. La corruzione può essere quantizzata dal numero di ufficiali del governo regionali perseguiti per pratiche corrotte. L'uso dell'indice di corruzione percepita (CPI), prodotto da Transparency International (1999), pur contenendo informazioni di grande valore, ha delle inconvenienze: il significato di corruzione è soggettivo e può variare da un Paese all'altro, in più i tipi di attività corrotte può essere sostanzialmente differente da paese a paese, rendendo così un possibile paragone difficile. D'altra parte l'utilizzo del numero di perseguiti ha i suoi lati negativi, si consideri ad esempio il fatto che in paesi corrotti di solito lo stesso sistema giudiziario è corrotto; si assume che tale problema sia sparso su tutto il territorio nazionale omogeneamente. Tuttavia in Italia, pur se il sistema legale è uguale in tutte le Regioni, il suo livello di legittimità non lo è. Secondo una ricerca di Hilary Putnam (1993), pur non essendo direttamente collegato alla corruzione, il livello di legittimità ha un impatto sull'efficienza dei governi regionali in Italia. Bisogna anche tenere conto del fatto che viene trascurata la corruzione non rilevata. Tuttavia i dati giudiziari rimangono le misure di corruzione possibili più efficaci. Nel capitolo terzo verranno identificate le diverse variabili che mettono in pratica le ipotesi considerate in questo capitolo per iniziare un'analisi quantitativa di verifica delle ipotesi.

CAPITOLO 3

ANALISI QUANTITATIVA DEL FENOMENO DI CORRUZIONE

3.1 VARIABILI QUANTITATIVE

L'analisi quantitativa si baserà su diverse variabli prese dal data-set provinciale del Rapporto attrattività e competitività dei territori italiani (2014). Le variabili verranno descritte in seguito.

L'indice di corruzione, cioè la variabile dipendente, verrà costruito attraverso 4 variabili giudiziarie. Le singole variabili saranno prima standardizzate per poter, in seguito, calcolare la media tra di loro.
Le variabili sono le seguenti:

1) Giustizia penale: usura
2) Giustizia penale: associazione per delinquere
3) Giustizia amministrativa: appalti pubblici di lavori, servizi e forniture
4) Giustizia amministrativa: autorizzazioni e concessioni

I dati provengono da ISTAT giustizia e sono precisamente il numero di delitti commessi da condannati con sentenza irrevocabile per ogni tipologia di crimine e sono standardizzati per la popolazione.

Per chiarezza espositiva e facilità di comparazione l'indice di corruzione è trasformato in maniera tale da rendere il valore minimo dell'indice di corruzione pari a zero, e il valore massimo pari a 100. La trasformazione è descritta dalla seguente relazione:

$$y^* = \frac{y-min}{MAX-min} \times 100,$$

dove y^* è l'indice di corruzione normalizzato e varia da 0 a 100, mentre y è l'indice di corruzione e varia da un valore minimo (min) a un valore massimo (MAX). Di seguito riporto una classifica delle province per il valore dell'indice:

PROVINCE	**INDICE DI CORRUZIONE (NORMALIZZATO)**
Bolzano	0
Trento	3
Forlì Cesena	5

Vercelli	6
Prato	6
Siena	6
Belluno	7
Treviso	7
Asti	9
Bologna	9
Fermo	9
Lodi	10
Mantova	10
Ferrara	10
Massa Carrara	10
Pisa	10
Pesaro e Urbino	10
Monza e della Brianza	11
Vicenza	11
Cuneo	12
Como	12
Padova	12
Venezia	12
Modena	12
Rimini	12
Torino	13
Verbano	13
Lecco	13
Arezzo	13
Ascoli Piceno	13
Novara	14
Verona	14
Pordenone	14
Udine	14
Firenze	14
Lucca	14
Sondrio	15
Piacenza	15
Ravenna	15
Bergamo	16
Varese	16
Carbonia-Iglesias	16
Medio Campidano	16
Ogliastra	16
Oristano	16
Olbia-Tempio	17
La Spezia	18
Biella	19
Brescia	19

Pavia	19
Alessandria	21
Genova	21
Milano	21
Imperia	22
Reggio Emilia	22
Pistoia	22
Ancona	22
Macerata	22
Sassari	22
Parma	23
Livorno	23
Cagliari	23
Nuoro	23
Gorizia	24
Bari	26
Trieste	27
Terni	28
Cremona	29
Teramo	29
Rovigo	30
Barletta-Andria-Trani	31
Grosseto	32
Savona	37
L'Aquila	38
Agrigento	39
Rieti	41
Benevento	41
Vibo Valentia	41
Matera	42
Catanzaro	43
Crotone	44
Lecce	45
Reggio di Calabria	45
Salerno	46
Taranto	46
Roma	47
Viterbo	47
Enna	47
Palermo	47
Avellino	48
Napoli	48
Foggia	49
Catania	49
Potenza	50
Cosenza	50

Brindisi	51
Latina	52
Caltanissetta	52
Messina	54
Frosinone	55
Siracusa	55
Trapani	55
Aosta	59
Campobasso	61
Caserta	62
Isernia	66
Chieti	68
Pescara	69
Perugia	77
Ragusa	100

In seguito saranno determinate le diverse variabili esplicative, tenendo conto di alcune delle ipotesi che derivano dalla letteratura, che potrebbero spiegare l'indice di corruzione. L'analisi è volta a trovare le cause del fenomeno di corruzione e il modo in cui queste si distribuiscono sul diversificato territorio nazionale italiano.

- Secondo l'ipotesi di Lipset (1960), la corruzione dipende positivamente dallo sviluppo economico e dal livello di istruzione. Per testare tale ipotesi si imposta un test di verifica sulle seguenti variabili a livello provinciale:

  1) Reddito disponibile delle famiglie consumatrici pro capite. (reddito)
     I dati provengono dall'appendice statistica al Rapporto 2015 di Unioncamere. Per dare un'idea della seguente distribuzione si consideri che in media il reddito disponibile delle famiglie consumatrici pro-capite è di 16355 euro annuali e mediamente vi è una variazione di 3463 euro, considerando che il range interquartile è di 5980 euro circa, si può concludere che all'interno del territorio italiano vi è una modesta diseguaglianza in termini di reddito a livello provinciale.

  2) PIL pro-capite. (PIL)
     I dati provengono dall'appendice statistica al Rapporto 2015 di Unioncamere. Il PIL pro-capite per provincia è naturalmente più alto rispetto al Reddito disponibile ma la sua distribuzione è simile in termini di variabilità, avendo un simile coefficiente di variazione (0.28 il PIL e 0.21 il Reddito).

3) Spesa pubblica per consumi finali per l'istruzione e la formazione (dato regionale): (spesa regione *PILprovincia/PILregione)/PIL provincia. (spesapubblicaist)
   I dati provengono da Istat- Noi Italia, e sono provincializzati come descritto.

4) Giovani che abbandonano prematuramente gli studi: popolazione 18-24 anni con al più la licenza media e che non frequenta altri corsi scolastici o svolge attività formative superiori ai 2 anni (per 100 persone della stessa età). (abbandono)
   I dati provengono da Istat - Banca dati di indicatori territoriali per le politiche di sviluppo - Istruzione e formazione, sono dati regionali, provincializzati. La media è di 20 ma il *range* interquartile è di 9.7, il che indica una grande differenza tra le province italiane.

5) Corsi di laurea del vecchio ordinamento per 100.000 abitanti in età >17 anni. (corsidilaureavecchio)
   I dati provengono da Istat- L'Italia in cifre. Questi dati non sono indice del livello di istruzione in una provincia in quanto all'interno del territorio italiano c'è un alto livello di mobilità, in maniera particolare per il grado di studi superiore. Vale la stessa considerazione per le variabili numero 6, 7, 8 e 9.

6) Corsi di dottorato per 100.000 abitanti in età >22 anni (corsidottorato)
   I dati provengono da Istat- L'Italia in cifre

7) Posti in aule per 100 studenti immatricolati e iscritti (postiinaula)
   I dati provengono da Istat- L'Italia in cifre

8) Docenti per 100 studenti immatricolati e iscritti (docentiognicento)
   I dati provengono da Istat- L'Italia in cifre

9) Studenti immatricolati e iscritti per docente (studentipersocente)
   I dati provengono da Istat- L'Italia in cifre

- Secondo la gamma di ipotesi che mette in relazione la corruzione alla presenza dello stato nell'economia si ritiene che all'aumentare della presenza dello Stato nell'economia la corruzione aumenti in quanto aumenta il campo di azione di pratiche corrotte. Tale ipotesi si può testare attraverso le seguenti variabili:

1) Finanziamenti iniziali dell'impresa per regione - Aiuti pubblici (valori % fatto 100 il totale delle nuove imprese). (finanziamentiinizio)
   I dati provengono da Istat - Le nuove attività imprenditoriali. In media gli aiuti consistono in 8,6 mila euro, ma anche in questo caso vi è una grande differenza tra minimo (1000 e 34100), quindi un alto range.

2) Incentivi alle imprese: Contributi agli investimenti alle imprese. (incentivi)
   I dati provengono da Istat - Le nuove attività imprenditoriali. Il modo in cui è distribuita tale variabile è simile alla variabile precedente. Un alto range caratterizza la variabilità all'interno del territorio italiano in termini di aiuti alle imprese.

- L'ipotesi di Alfredo Del Monte e Erasmo Papagni (2007), secondo cui il livello di sviluppo economico influenza il livello di corruzione, viene verificata con la seguente variabile:

  1) Livello di sviluppo economico: Variazione percentuale del PIL. (crescita)
     I dati provengono dall'appendice statistica al Rapporto 2015 di Unioncamere.

- Verrà anche analizzata un'ulteriore ipotesi, non tratta direttamente dalla letteratura, ma formulata tenendo presenti letteratura e considerazioni personali. L'inefficienza del tessuto imprenditoriale potrebbe essere causa di un aumento del livello di corruzione, dato che produce mancanza di trasparenza, con conseguente attrazione del reato di corruzione, che trova di base un terreno oscurato. L'inefficienza riflette difficoltà economiche e mancanza di organizzazione all'interno della pubblica amministrazione, rendendo vulnerabile il tessuto imprenditoriale e dunque più facilmente corruttibile. Tale ipotesi può essere testata considerando le seguenti variabili che riflettono l'inefficienza del tessuto imprenditoriale:

  1) Starting a business: Time (days) (starttime)
     I dati provengono da "Doing Business in Italy 2013" e corrispondono ai giorni necessari per aprire una nuova attività imprenditoriale.

  2) Starting a business: Cost (startcost)
     I dati provengono da "Doing Business in Italy 2013" e corrispondono al costo medio necessario per aprire una nuova attività imprenditoriale.

3) Difficoltà nell'iniziare una nuova attività imprenditoriale: reperire finanziamenti molto /in parte. (startfinanz)
   I dati provengono da Istat - Le nuove attività imprenditoriali. I dati sono basati sulle di un questionario relativo alle attività imprenditoriali. Le prossime due variabili sono misurate nella stessa maniera.

4) Difficoltà nell'iniziare una nuova attività imprenditoriale: trovare personale qualificato molto / in parte. (startpersqaulif)
   I dati provengono da Istat - Le nuove attività imprenditoriali.

5) Difficoltà nell'iniziare una nuova attività imprenditoriale: aspetti giuridici amministrativi molto /in parte. (startgiuramm)
   I dati provengono da Istat - Le nuove attività imprenditoriali.

6) Durata dei procedimenti in materia di lavoro subordinato 2006. (duratagiust)
   I dati provengono da Istat - Sistema Informativo Territoriale. Espressa in termini di giorni, la media è di 1100 circa, con un range altissimo, in quanto il minimo è di 107, il massimo di 5942. La variabilità all'interno del campione è altissima, tanto da avere una deviazione standard di 971 giorni circa.

Per l'analisi si utilizzerà il software statistico GRETL.

3.2 MODELLO DEI MINIMI QUADRATI

Prendendo tutte le variabili in considerazione, si formula il modello di regressione lineare che verrà stimato con il metodo OLS, i *p*-value significativi saranno quelli contrassegnati con * se <0.10; ** se <0.05; *** se <0.01.

Modello 1: OLS, usando le osservazioni 1-110 (n = 51)
Sono state scartate osservazioni mancanti o incomplete: 59
Variabile dipendente: CORRUZIONE

Media var. dipendente	0,057065		SQM var. dipendente	0,653642
Somma quadr. residui	5,938767		E.S. della regressione	0,430797
R-quadro	0,721999		R-quadro corretto	0,565623

F(18, 32)	4,617075		P-value(F)	0,000084
Log-verosimiglianza	–17,53260		Criterio di Akaike	73,06521
Criterio di Schwarz	109,7699		Hannan-Quinn	87,09115

	Coefficiente	*Errore Std.*	*rapporto t*	*p-value*	
const	–3,0634	1,77627	-1,7246	0,09424	*
Reddito	3,06994e-05	6,71986e-05	0,4568	0,65087	
PIL	–7,53463e-05	4,6533e-05	-1,6192	0,11522	
spesapubblicaist	0,00060578	0,000531243	1,1403	0,26262	
abbandono	–0,0170067	0,0138101	-1,2315	0,22712	
corsidilaureavecchio	–0,0504467	0,0221405	-2,2785	0,02951	**
corsidottorato	–0,0202068	0,0164576	-1,2278	0,22848	
postiinaula	0,00914675	0,0139065	0,6577	0,51541	
docentiognicento	–0,00328651	0,0070739	-0,4646	0,64537	
studenitperdocente	–0,000327549	0,000524935	-0,6240	0,53706	
finanziamentiinizio	–0,0322196	0,021949	-1,4679	0,15188	
incentivi	–1,25435e-06	1,1477e-06	-1,0929	0,28258	
crescita	0,335415	0,14605	2,2966	0,02834	**
starttime	–0,0385654	0,0492317	-0,7833	0,43918	
startcost	0,010075	0,0736958	0,1367	0,89212	
startfinanz	0,0129088	0,00983542	1,3125	0,19869	
startpersqualif	0,0635357	0,0229508	2,7683	0,00929	***
startgiuramm	0,0118713	0,0475858	0,2495	0,80459	
duratagiust	0,000100206	9,05395e-05	1,1068	0,27665	

Dall'analisi di tutte le variabili, notiamo che molte non sono significative. Si cerca un metodo per eliminare le variabili non significative e giungere a un modello con solo variabili significative almeno al 5%.
Per la selezione delle variabili si agirà con il metodo "try backward". Il criterio adottato è quello di eliminare una variabile alla volta e impostare

nuovamente l'analisi dei minimi quadrati (OLS), sino a raggiungere un modello in cui tutte le variabili siano significative. Nel primo modello, ad esempio si imposta una nuova analisi senza la variabile "Reddito", in quanto ha il *p*-value più alto (0,65). Si consideri che per ogni ipotesi tratta dalla letteratura, si hanno diverse variabili o indicatori. Qualora rimanga un'unica variabile riferita a una determinata ipotesi e questa sia quella con il *p*-value più alto, non verrà scartata; al suo posto verrà scartata la variabile con il secondo *p*-value più alto. Questo, per cercare di mantenere il più possibile sotto controllo ogni ipotesi. Bisogna tenere presente che se una variabile non è significativa non si può concludere che non vi sia nesso di causalità tra questa e l'indice di corruzione. Infatti, talvolta, la mancata significatività di una variabile x_i sull'indice di corruzione può essere data dal fatto che un'altra variabile x_j spiega di per sé l'effetto che x_i ha sulla corruzione.
In seguito all'applicazione del metodo sopra descritto, dopo 17 modelli, si raggiunge un modello di regressione multipla che presenta le seguenti caratteristiche:

Modello 18: OLS, usando le osservazioni 1-110
Variabile dipendente: CORRUZIONE

Media var. dipendente	0,000000		SQM var. dipendente	0,657332
Somma quadr. residui	21,31220		E.S. della regressione	0,452686
R-quadro	0,547485		R-quadro corretto	0,525730
F(5, 104)	25,16536		P-value(F)	1,50e-16
Log-verosimiglianza	–65,81720		Criterio di Akaike	143,6344
Criterio di Schwarz	159,8373		Hannan-Quinn	150,2064

	Coefficiente	*Errore Std.*	*rapporto t*	*p-value*	
const	–2,77663	0,6533	-4,2502	0,00005	***
PIL	–3,01134e-05	1,22343e-05	-2,4614	0,01548	**
finanziamentiinizio	–0,0240359	0,00915003	-2,6269	0,00992	***
spesapubblicaist	0,000510677	0,0001980 98	2,5779	0,01134	**
starttime	–0,037927	0,0202168	-1,8760	0,06346	*

	1				
startpersqualif	0,0586954	0,00988215	5,9395	<0,00001	***

Prendendo come livello di significatività *p*-value con valori minori di 0.05, la variabile "start-time" viene scartata, il modello che si raggiunge è il seguente:

Modello 19: OLS, usando le osservazioni 1-110
Variabile dipendente: CORRUZIONE

Media var. dipendente	0,000000		SQM var. dipendente	0,657332
Somma quadr. residui	22,03342		E.S. della regressione	0,458085
R-quadro	0,532172		R-quadro corretto	0,514350
F(4, 105)	29,86035		P-value(F)	1,38e-16
Log-verosimiglianza	–67,64763		Criterio di Akaike	145,2953
Criterio di Schwarz	158,7977		Hannan-Quinn	150,7719

	Coefficiente	*Errore Std.*	*rapporto t*	*p-value*	
const	–2,83454	0,660353	-4,2925	0,00004	***
PIL	–2,53142e-05	1,21065e-05	-2,0910	0,03895	**
finanziamentiinizio	–0,0199536	0,0089935	-2,2187	0,02866	**
spesapubblicaist	0,000456254	0,000198299	2,3008	0,02338	**
startpersqualif	0,0529805	0,00951301	5,5693	<0,00001	***

3.3 LE VARIABILI SIGNIFICATIVE

Il modello finale di regressione multipla risulta formato da quattro variabili significative:

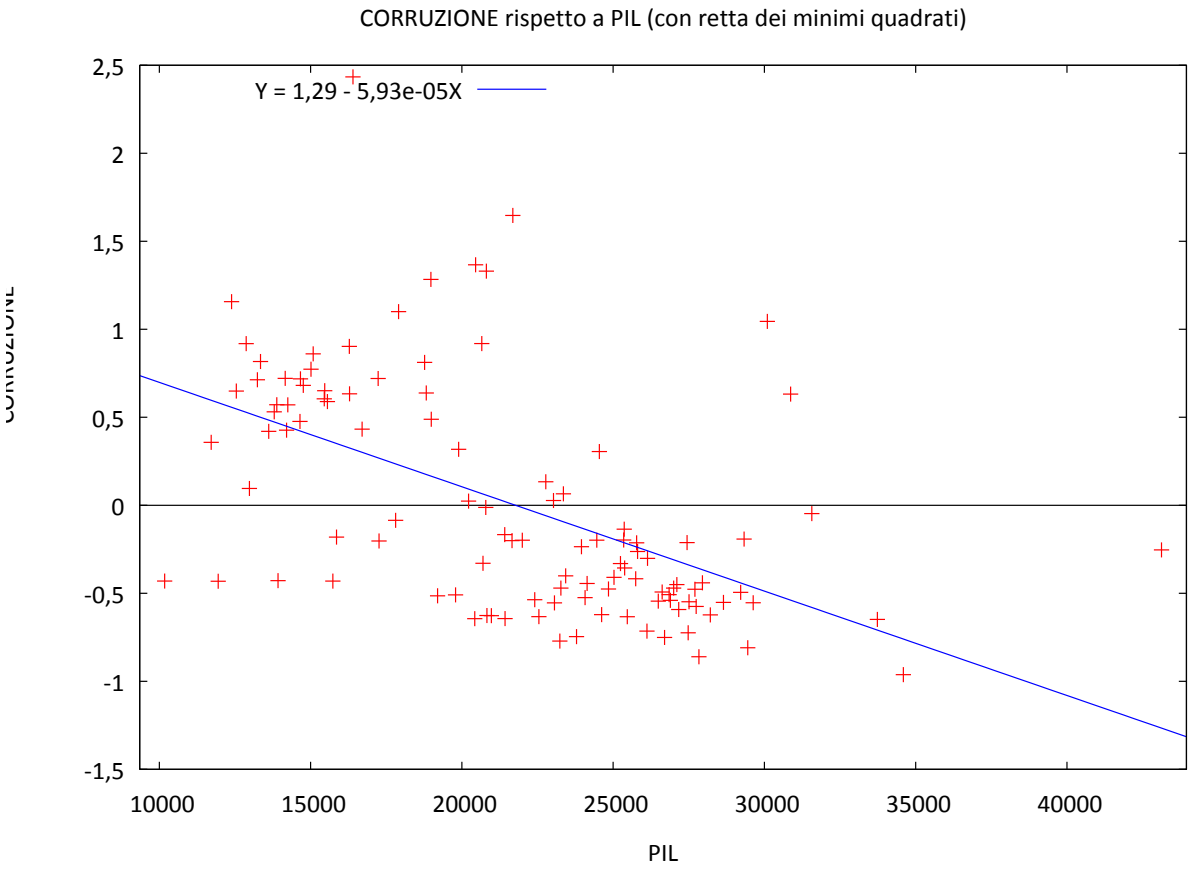
- PIL pro-capite, che ha un coefficiente negativo.
- Finanziamenti iniziali dell'impresa per regione - Aiuti pubblici (valori % fatto 100 il totale delle nuove imprese), che ha un coefficiente negativo.

- Spesa pubblica per consumi finali per l'istruzione e la formazione (dato regionale): (spesa regione*PIL provincia/PIL regione)*PIL provincia, che ha un coefficiente positivo.
- Difficoltà nell'iniziare una nuova attività imprenditoriale: trovare personale qualificato molto /in parte, che ha un coefficiente positivo.

Prendendo in considerazione le variabili all'interno del modello di regressione multipla, il modo in cui influiscono sull'indice di corruzione è descritto dal coefficiente di regressione; questo risulta conforme alla letteratura eccetto per quanto riguarda la variabile "Spesa pubblica per consumi finali per l'istruzione e la formazione" e la variabile "finanziamenti iniziali".
Prendendo in considerazione le variabili del modello di regressione multipla singolarmente, ognuna da sola ha una relazione lineare con l'indice di corruzione, che può essere visualizzata graficamente. In tale analisi grafica, la variabile "finanziamenti iniziali" risulta avere un coefficiente positivo, questa volta coerente con la letteratura a riguardo. Questo ci dice che il segno negativo del suo coefficiente è dovuto alla collinearità con le altre variabili. Nell'analisi grafica, la "spesa pubblica per consumi finali per l'istruzione e la formazione" mantiene un coefficiente positivo, non coerente con la letteratura. In seguito commenteremo ulteriormente tale risultato. Per analizzare con dettaglio ogni relazione si imposta un modello OLS per ogni variabile. Considerando il *p*-value, ognuna risulta significativa, ma i valori degli indici di determinazione, r^2, sono bassi, a testimonianza del fatto che un modello lineare semplice non è in grado di provvedere la stessa quantità di informazione rispetto a quelle ricavabili dal modello lineare multiplo. Ciò è valido soprattutto per un fenomeno così complesso come la corruzione.

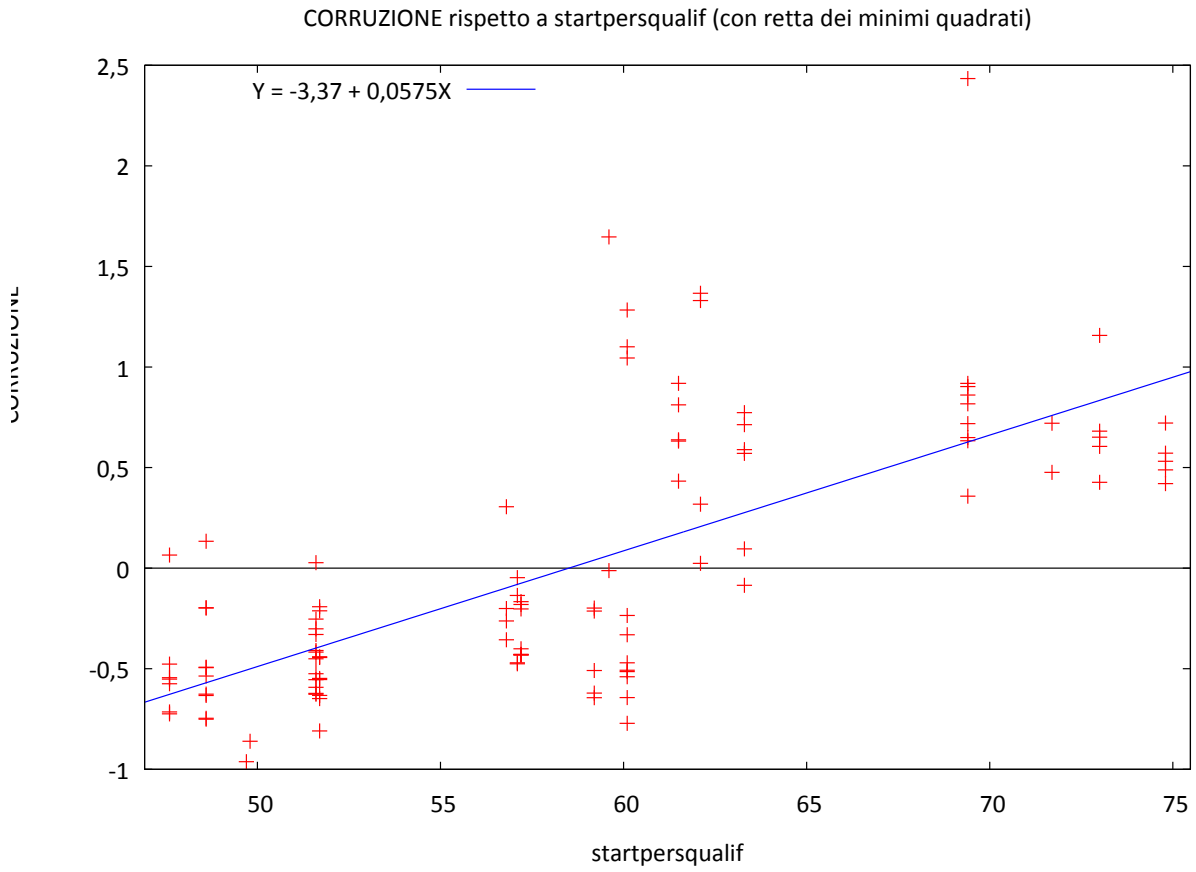
3.4 ANALISI GRAFICA

Di seguito vi sono i grafici a dispersione delle variabili significative prese singolarmente e l'indice di corruzione con i relativi indici di determinazione e il corrispettivo *p*-value.



x_1: PIL procapite
y: Indice di corruzione
r^2 = 0,292645
p-value = 1,05e-09

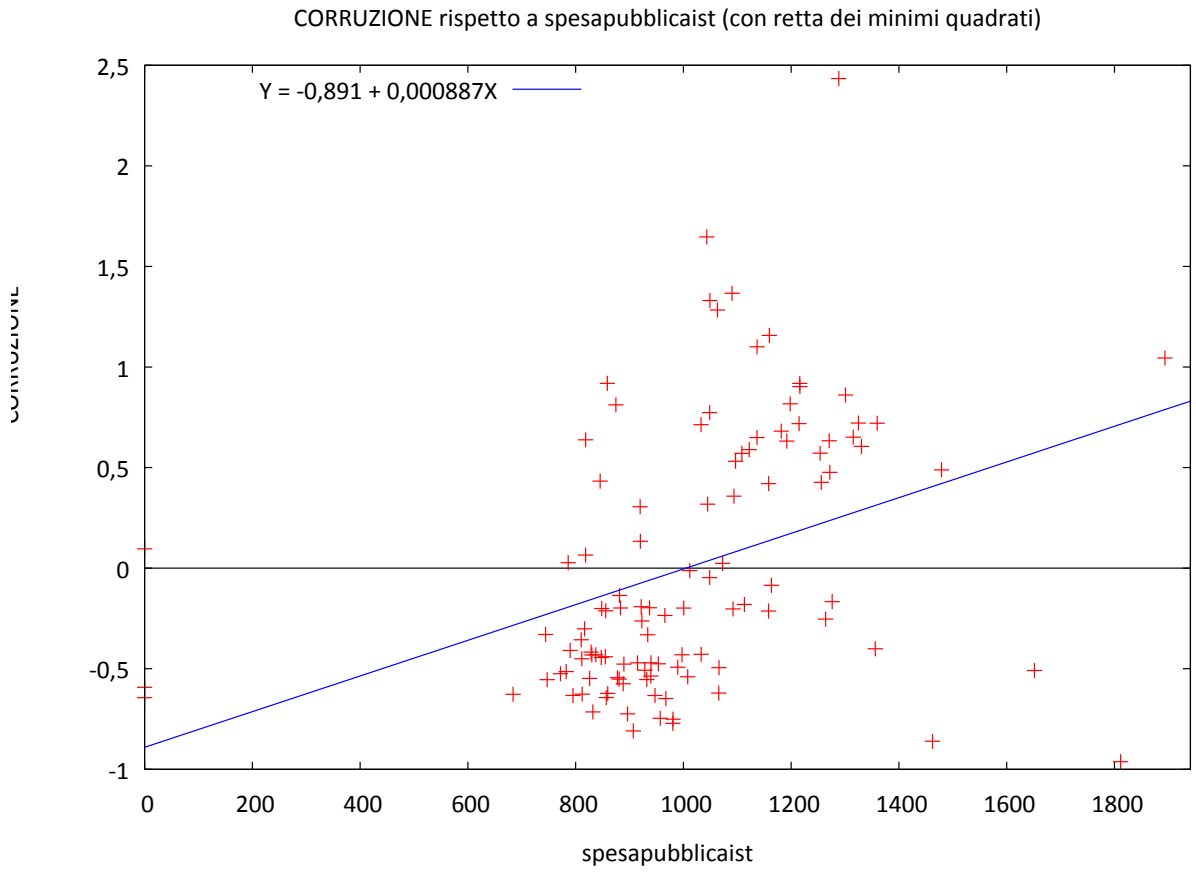
L'indice di determinazione è basso, in quanto, come si è detto in precedenza, di per sé il PIL pro-capite non riesce a spiegare il fenomeno di corruzione in toto. Tuttavia il PIL pro-capite è significativo dato il valore del p-value. Il verso della relazione è conforme alla letteratura al riguardo: al crescere della ricchezza cresce il controllo sulla pubblica amministrazione e diminuisce la corruzione.



x_2: Finanziamenti iniziali dell'impresa per regione - Aiuti pubblici (valori % fatto 100 il totale delle nuove imprese)
y: Indice di corruzione

r^2 = 0,224628
p-value = 1,70e-07

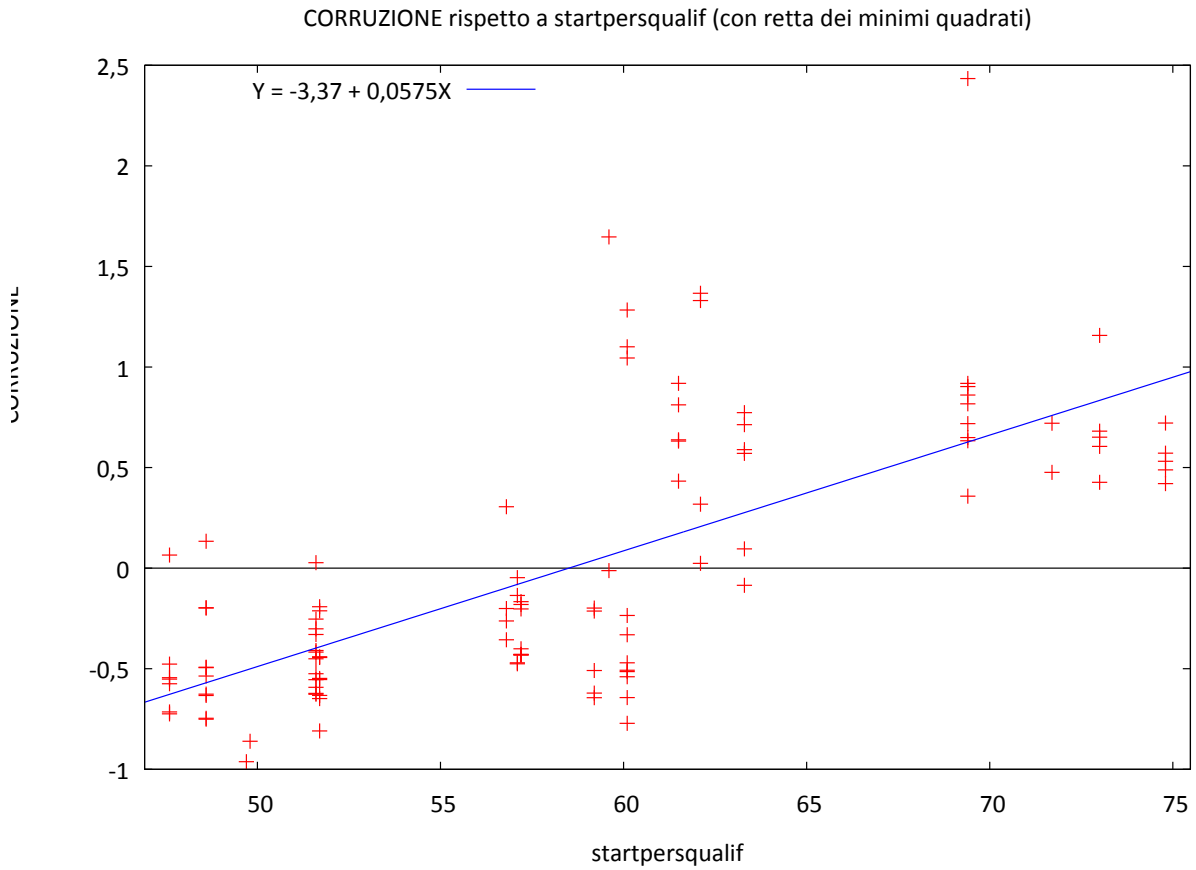
Anche in questo caso l'indice di determinazione ha un valore basso, in quanto non è sufficiente la variabile x_2 per spiegare l'indice di corruzione. Tuttavia risulta significativa e coerente con la letteratura, in quanto il verso della relazione tra x_2 e y è positivo.



x_3: Spesa pubblica per consumi finali per l'istruzione e la formazione (dato regionale): (spesa regione in % del PIL*PIL regione)*peso PIL provincia.
y: Indice di Corruzione

r^2 = 0,136007
p-value = 7,36e-05

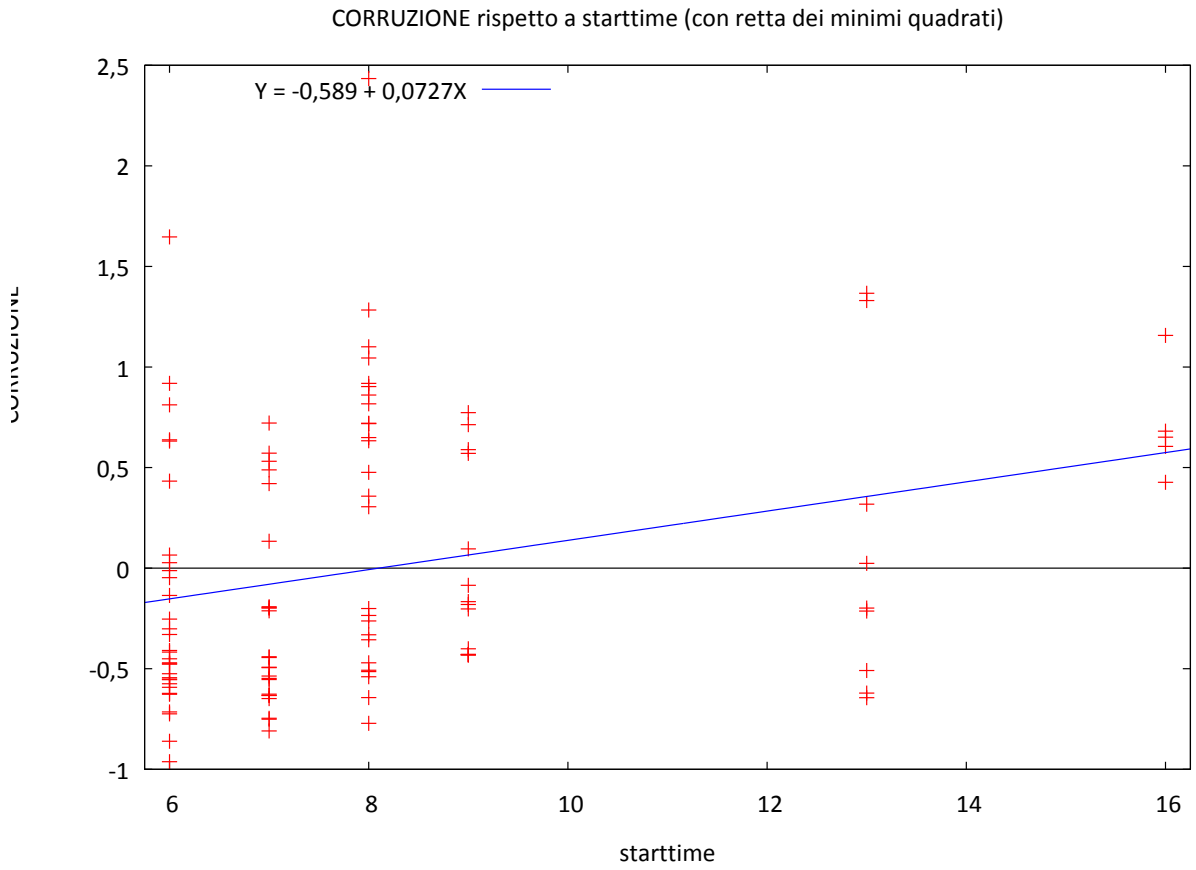
La spesa pubblica per l'istruzione risulta essere connotata da un indice di determinazione basso. Tuttavia, il valore del *p*-value mostra la sua significatività. Il verso della relazione, però, non risulta conforme alla letteratura. Al crescere della spesa per l'istruzione dovrebbe crescere il livello di istruzione che porta ad un controllo attivo sull'operato politico, il che disincentiva pratiche corrotte. Questo non porta direttamente ad escludere la veridicità di tale ipotesi in quanto tale risultato può essere dovuto al fatto che la spesa pubblica per consumi non sia un indice adatto a misurare il livello d'istruzione di una provincia, in quanto c'è la possibilità che province povere e dunque con un alto tasso di corruzione (come si è dimostrato dalla relazione tra PIL e indice di corruzione) abbiano una spesa per istruzione, in proporzione al PIL, maggiore.



x_4: Difficoltà nell'iniziare una nuova attività imprenditoriale: trovare personale qualificato molto /in parte
y: Indice di Corruzione

r^2 = 0,495617
p-value = 9,59e-18

Tra tutte le variabili appartenenti al modello di regressione multipla trovato, la difficoltà nel trovare personale qualificato è la variabile con il più alto indice di determinazione, presa singolarmente. Risulta significativa e coerente con l'ipotesi dell'inefficienza formulata nel capitolo secondo e riprodotta anche nella variabile x_5.



x_5: Doing Buisness: Start Time (days)
y: Indice di Corruzione

$r^2 = 0{,}079730$
p-value = 0,0028

Tra le varie variabili, la variabile "giorni necessari per l'inizio di una nuova attività imprenditoriale" è quella con l'indice di determinazione più basso, il suo livello di significatività all'interno del modello lineare multiplo instaurato, ha portato ad escluderla essendo il p-value maggiore di 0,05. Risulta comunque interessante analizzare la sua incidenza sul livello di corruzione. Essa risulta essere significativa, presa singolarmente e coerente con l'ipotesi sulla inefficienza, in quanto l'inefficienza porta a rallentare i tempi per l'inizio di una nuova attività; questa variabile, dunque, rispecchia inefficienza, la quale porta maggiore corruzione.

CONCLUSIONE

Il modello di regressione lineare multipla trovato risulta conforme ad alcune delle ipotesi prese dalla letteratura e all'ipotesi formulata in seguito a riflessioni personali. Prendendo in considerazione l'ipotesi di Lipset (1960), la variabile x_1, ovvero PIL pro-capite, indica che lo sviluppo economico delle province italiane ha una relazione negativa con l'indice di corruzione. Infatti, come si può vedere anche graficamente, al crescere di tale variabile l'indice di corruzione diminuisce. In effetti, secondo l'ipotesi formulata da Lipset (1960), un elevato livello economico della popolazione garantisce un controllo sociale sul settore pubblico molto forte, tanto da disincentivare pratiche corrotte, grazie all'uso della partecipazione politica attiva durante le elezioni. Pratiche corrotte vengono disincentivate grazie alla pressione elettorale. Di per sé non è sufficiente considerare solamente questo aspetto come unica causa di un aumento o di una diminuzione del livello di corruzione. Infatti l'analisi effettuata prende in considerazione altre ipotesi, che effettivamente hanno un effetto sull'indice di corruzione. Si prenda in considerazione l'ipotesi che mette in relazione la presenza dello stato nell'economia con il livello di corruzione. La presenza dello stato nell'economia risulta rilevante se si considera la variabile degli aiuti pubblici alle imprese x_2. In effetti, all'aumentare di x_2, aumenta la presenza dello Stato nell'economia e ciò incrementa le potenziali situazioni in cui si possono instaurare pratiche corrotte. Considerando graficamente la relazione tra x_2 e l'indice di corruzione y, essi sono positivamente correlati. Così come x_1 e x_3, anche la variabile x_2 può spiegare solo in parte l'aumento o la diminuzione dell'indice di corruzione. Continuando l'analisi, l'ipotesi di Del Monte e Papagni (2007) non risulta essere verificata. Come si è spiegato in precedenza, ciò non vuol dire che lo sviluppo economico non abbia effetto sulla corruzione, ma può spiegare come lo sviluppo economico sia rappresentato da altre variabili presenti nel modello e, dunque, la sua aggiunta nel modello di regressione lineare risulta essere non influente. A differenza di questa conclusione, l'ipotesi secondo la quale l'inefficienza avrebbe un effetto positivo sull'indice di corruzione, formulata in seguito a considerazioni personali, può essere dimostrata dall'influenza delle seguenti variabili sull'indice di corruzione: la difficoltà nel trovare personale qualificato per iniziare una nuova attività imprenditoriale e il tempo necessario per iniziare una nuova attività imprenditoriale, rispettivamente x_4 e x_5. La difficoltà nell'intraprendere una nuova attività imprenditoriale riflette l'inefficienza nel settore economico privato. Tale inefficienza comporta la mancanza di trasparenza e l'aumento di pratiche corrotte, che trovano terreno fertile appunto in ambiti poco trasparenti. La corruzione a

sua volta alimenta l'inefficienza, creando un circolo vizioso che viene riflesso dal forte legame tra le variabili considerate: x_4 e y e in parte, x_5 e y.
Le variabili x_1, x_2, x_3 e x_4, prese in considerazione insieme sono le variabili indipendenti del modello di regressione lineare multiplo, in cui la variabile dipendente è l'indice di corruzione y. Tali variabili riescono a spiegare parte del fenomeno di corruzione, portando a un indice di determinazione r^2 di 0,53.

APPENDICE

Le variabili utilizzate, per la ricerca quantitativa sulle cause della corruzione, provengono dal database provinciale del rapporto attrattività e competitività dei territori italiani (2014). Ringrazio la dott.ssa Francesca Sica per aver concesso l'utilizzo dei dati del rapporto per la seguente Tesi e per l'aiuto offerto durante la stesura. Il database provinciale è stato da me predisposto per una collaborazione con Confindustria-CeFOP (Centro Studi di Economia della Formazione e delle Professioni). I dati non reperibili a livello provinciale sono dati regionali provincializzati, tenendo conto della popolazione delle province e altri dati provinciali. Il rapporto attrattività e competitività dei territori italiani (2014) ha come finalità quella di "descrivere i risultati di una ricerca (…) volta a elaborare una nuova misura di attrattività territoriale atta a catturare il potenziale attrattivo, in senso produttivo e residenziale, dei territori italiani, regioni e province, compensando i limiti degli indicatori tradizionali di attrattività". Pur avendo finalità diverse, la varietà dei dati presenti nel database ha consentito l'avvio della ricerca per la presente tesi.

Di seguito una descrizione di ogni variabile: la descrizione analizza la distribuzione di ogni variabile considerando media, mediana, valore minimo, valore massimo, variabilità (deviazione standard e coefficiente di variazione), asimmetria, curtosi, percentili e range interquartile.

Statistiche descrittive, usando le osservazioni 1 - 110

Variabile	Media	Mediana	Minimo	Massimo
PIL	21778,0	22202,5	10176,1	43125,8
crescita	-0,503132	-0,430411	-3,41410	1,44619
incentivi	58688,1	38293,3	1644,00	507710,
istruzione	92,1013	91,8750	71,9000	114,250
Reddito	16355,1	16643,2	10534,8	26733,3
finanziamentiinizio	8,62273	5,10000	1,00000	34,1000
tassodinatalitaimp	0,109667	0,0934171	0,0107757	0,788222
impreseattsupop	76,3250	74,6327	23,5325	307,753
spesapubblicaist	1004,56	973,886	0,00000	1893,69
abbandono	19,3551	18,5000	7,50000	41,3000
corsidilaureavecchio	4,42671	3,67000	0,00000	42,3400

corsidottorato	11,1337	12,5000	0,00000	27,6600
postiinaula	32,8920	33,5350	7,13000	52,2700
docentiognicento	24,6168	22,0600	2,36000	97,6800
studenitperdocente	196,145	63,2800	30,3800	6166,67
starttime	8,10000	7,00000	6,00000	16,0000
startcost	14,7755	14,5000	12,2000	16,8000
startfinanz	55,1000	55,0000	0,00000	97,0000
startpersqualif	58,5036	57,2000	47,6000	74,8000
startgiuramm	29,8836	29,5000	24,1000	39,0000
duratagiust	1099,84	745,000	107,000	5942,00
Variabile	**Dev. Std.**	**Coeff. di variazione**	**Asimmetria**	**Curtosi**
PIL	5996,73	0,275358	0,274900	0,117523
crescita	0,918675	1,82591	-0,578068	0,276568
incentivi	78387,6	1,33567	3,76281	16,4246
istruzione	7,63815	0,0829321	0,0264517	0,542862
Reddito	3462,75	0,211723	0,186876	-0,585617
finanziamentiinizio	8,27951	0,960196	1,34823	1,65669
tassodinatalitaimp	0,0986554	0,899588	4,59168	24,5982
impreseattsupop	28,6522	0,375397	5,03075	37,3358
spesapubblicaist	273,436	0,272195	-0,522520	4,52204
abbandono	6,28663	0,324804	0,662310	0,444688
corsidilaureavecchio	6,22933	1,40721	4,43564	22,5725
corsidottorato	5,86058	0,526382	-0,401196	0,325428
postiinaula	8,40437	0,255514	-0,685849	0,856517
docentiognicento	13,5405	0,550052	2,92326	14,1782
studenitperdocente	745,216	3,79931	7,55476	57,6559
starttime	2,55203	0,315065	1,77768	2,52884
startcost	1,47519	0,0998405	-0,116461	-1,22015
startfinanz	13,9448	0,253082	-1,20103	5,26090
startpersqualif	8,04282	0,137476	0,481507	-0,745794
startgiuramm	3,44534	0,115292	0,595703	0,608644
duratagiust	971,291	0,883122	2,61610	7,53507

Variabile	**5% Perc.**	**95% Perc.**	**Range interquartile**	**Osservazioni mancanti**
PIL	12724,1	30443,1	9940,15	0
crescita	-2,16629	0,757925	1,29387	0
incentivi	7030,88	167096,	49206,3	0
istruzione	78,2635	104,562	10,5225	0
Reddito	11177,7	21578,8	5981,82	3
finanziamentiinizio	1,00000	25,8500	11,2000	0
tassodinatalitaimp	0,0210595	0,277170	0,0252935	5
impreseattsupop	49,9713	110,932	21,2451	0
spesapubblicaist	745,815	1405,88	301,415	0
abbandono	11,2000	29,6200	9,70000	3
corsidilaureavecchio	0,00000	9,59800	4,40250	40
corsidottorato	0,00000	18,1700	7,14000	56
postiinaula	18,2445	45,3785	10,7100	40
docentiognicento	7,41200	40,6860	11,6100	57
studenitperdocente	34,5405	579,928	41,5300	40
starttime	6,00000	14,3500	3,00000	0
startcost	12,2000	16,8000	2,50000	0
startfinanz	39,5500	75,0000	15,2500	0
startpersqualif	47,6000	73,8100	10,8000	0
startgiuramm	24,1000	36,9100	3,52500	0
duratagiust	297,900	3491,00	582,000	5

BIBLIOGRAFIA

Agresti, A. Finlay, B. (2009), Statistical Methods for the Social Sciences, *Fourth Edition, Pearson.*

Becker, G. (1968). Crime and punishment: an economic approach *Journal of Political Economy 76.*

Cottrell, A. Lucchetti, R. (2014), Gretl User's Guide, *Gnu Regression, Econometrics and Time-series Library.*

De Giovanni, L. G.M. Sica, F. (2014) Parte Prima attrattività e competitività dei territori italiani i - la metodologia di misura; II – *I risultati: le dimensioni dell'attrattività territoriale*, Rivista di Politica Economica RPE TERRITORIAL (ottobre/dicembre) Anno CIII-SERIEIII, Fascicolo X-XII.

Del Monte, A. Papagni, E. (2007), The determinants of corruption in Italy: Regional panel data analysis. *European Journal of Political Economy* 23(2): 379-396.

G.M. Sica, F. (2015) Basic versus Innovation. *Verifica dell'esistenza e dell'intensità del legame tra Istituzioni e Capacità tecnologica: un 'analisi cross-countries*.

Fiorino, N. Galli E. (2010). An analysis of the determinants of corruption: Evidence from the Italian regions*. Working paper 171* (September).

Freedman, D. Pisani, R. Purves, R. (2007), Statistics, *Fourth Edition, W.W. Norton & Company*

Greco Eval RC- I/II Rep 2011, 1E.

Lijphart, A. (1999). Patterns of democracy, government forms and performance in thirty six countries. Yale University Press. New Haven

Lipset, S. (1960). *Political Man: The Social Bases of Politics*. Doubleday, Garden City, NY.

Ministero per la semplificazione e la pubblica amministrazione (2012). Rapporto della Commissione per lo studio e l'elaborazione di proposte in tema di trasparenza e prevenzione della corruzione nella pubblica amministrazione *http://www.funzionepubblica.gov.it/*

Pope, J. (1999) The Need for, and Role of, an Independent Anti-Corruption Agency. *Transparency International Working Paper*

Putnam R. (1993). Making Democracy Work: Civic Traditions in Modern Italy, Princeton, N.J.: *Princeton University Press*

Rose-Ackerman, S. (1975). The economics of corruption. *Journal of Public Economics* 4, 187– 203 (February)

Wolfensohn J. (1997). Helping Countries Combat Corruption: The Role of the World Bank 8-17 (September), The International Bank for Reconstruction and Development / the World Bank.