Sondaggi Elettorali

Studio sull'Affidabilità Statistica degli Istituti Demoscopici e delle loro Distorsioni Sistematiche

Tommaso Menghini

Università degli Studi di Milano-Bicocca

Cos'è un sondaggio elettorale I

- I sondaggi elettorali sono una delle manifestazioni più comuni della statistica nella vita di tutti i giorni.
- Ma l'**affidabilità** legata a questi ultimi è stata messa in discussione.
- Costruire un sondaggio d'opinione comporta **contattare** un certo numero di persone e **porre** una serie di domande, tra le quali anche chi si intende votare nella prossima elezione.
- Ci sono diverse ragioni per cui questo processo possa andare storto.
- **Problema chiave**. Sono presenti errori di campionamento e non. Ci si riferisce a questi come **total survey error**.
- L'**errore di campionamento** è il più gestibile. Di maggiore importanza, invece, sono i **nonsampling errors**.

Cos'è un sondaggio elettorale II

- I **nonsampling errors** non possono essere ridotti semplicemente conducendo più sondaggi, oppure contattando più persone.
- **Esempio**. L'**errore di non risposta** si verifica quando un intervistato, sostenitore di un partito che sta andando male nei sondaggi preferisce non rispondere. Questo meccanismo è alla base della **non risposta differenziale**.
- I sondaggisti minimizzano l'errore aggiustando i loro dati per differenze note tra il loro campione grezzo e la popolazione. Utilizzano tecniche di **post-stratificazione**.
- **Esempio**. Ad ogni intervistato viene associato un **peso numerico**, definito in modo tale che le **distribuzioni pesate** di variabili demografiche delle unità del campione grezzo corrispondano alle **distribuzioni marginali** riferite alla popolazione target.

Cos'è un sondaggio elettorale III

- **Problema**. Fallendo nel correggere per queste diverse probabilità d'inclusione si rischia di **distorcere** pesantemente le stime legate al campione ottenuto.
- Un sondaggio elettorale può essere visto come un'individualità, ma anche come parte di un contesto più ampio.
- L'**aggregazione** permette di combinare le stime di diversi sondaggi al fine di produrne una singola. Questa strategia è anche chiamata comunemente **Poll of Polls**.
- **Vantaggio**. **(presunto)** La media dei risultati provenienti da diversi istituti restituisce una stima più affidabile evitando di fondarsi sul dato potenzialmente instabile di un singolo sondaggio.

Origine dei dati I

- I dati sono stati resi disponibili dall'azienda **YouTrend**. Sondaggi elettorali condotti in Italia nell'intervallo temporale che inizia nel 2018 e finisce nel 2024.
- Per ogni sondaggio si ha il nome dell'istituto demoscopico che lo ha realizzato, la data, la numerosità campionaria e la stima delle intenzioni di voto per una serie di partiti.
- In questa finestra temporale si sono tenute quattro elezioni: le elezioni politiche del 2018 e del 2022, le elezioni europee del 2019 e del 2024.
- **Idea**. Valutare l'**affidabilità** associata ad ogni istituto demoscopico in base al confronto delle proprie previsioni con i risultati delle elezioni corrispondenti.

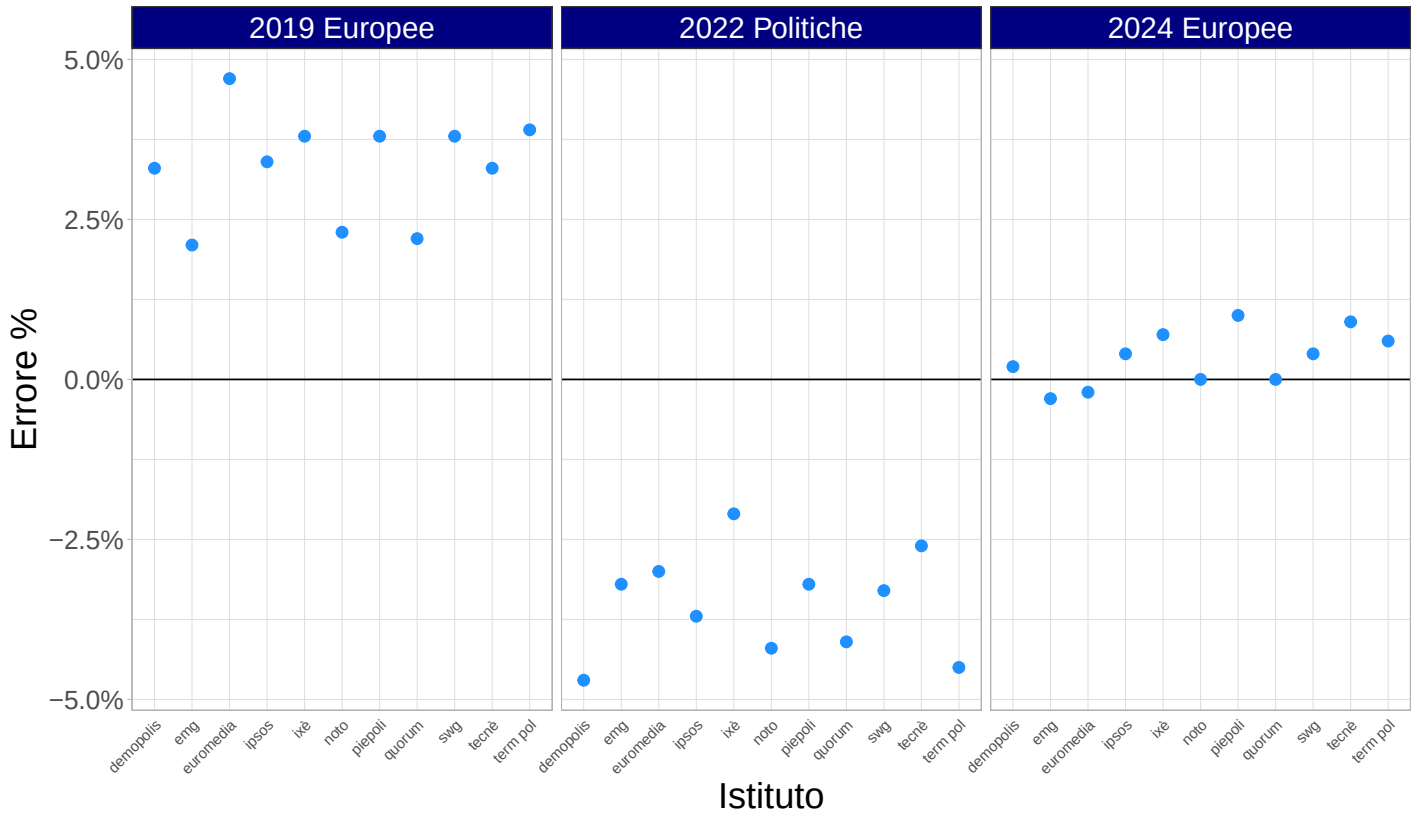
Riferimento principale

Shirani-Mehr, H., Rothschild, D., Goel, S., Gelman, A. (2018). *Disentangling Bias and Variance in Election Polls*. *Journal of the American Statistical Association*.

Origine dei dati II

- **Problema**. **(partiti)** La volatilità della realtà politica italiana rende difficile mantenere un insieme omogeneo di soggetti politici lungo tutto il periodo d'analisi.
- **Soluzione**. Si lavora su un insieme stabile e comparabile di soggetti politici: Fratelli d'Italia, Forza Italia, Lega, Movimento 5 Stelle e Partito Democratico.
- **Problema**. **(agenzie)** Non tutte le agenzie dispongono di una serie storica estesa e regolare a sufficienza.
- É necessario costruire un **confronto bilanciato**, in cui vi sia lo stesso numero di combinazioni elezione-partito per ogni agenzia.
- **Soluzione**. Si selezionano quelle agenzie che presentano almeno una rilevazione precedente alle Europee del 2019 e del 2024 e alle Politiche del 2022 e che garantiscono una copertura temporale sufficiente.

Intuizione: intenzioni di voto e risultati elettorali



- Differenza tra il risultato delle elezioni corrispondenti e le intenzioni di voto stimate dall'ultimo sondaggio disponibile per agenzia per la Lega.
- **Intuizione**. Non esistono differenze significative tra le undici agenzie analizzate in termini di affidabilità.

La variabile risposta

- L'uso diretto della differenza tra il risultato elettorale e la percentuale stimata risulta non idonea, in quanto non comparabile tra partiti diversi.
- Sia dunque $i = 1, ..., 11$ l'indice legato all'agenzia, $k = 1, ..., 5$ l'indice per il partito e $j = 1, 2, 3$ l'indice associato al turno elettorale. La **variabile risposta** viene definita come:

$$y_{ikj} = \frac{V_{kj} - \widehat{V}_{ikj}}{\widehat{V}_{ikj}}$$

- Questa formulazione della distorsione è l'**errore relativo**.
- Il dataset utilizzato per l'analisi è dunque composto dalle tre variabili categoriche indipendenti `Istituto`, `Partito` e `Elezione`.
- In totale si hanno a disposizione 165 osservazioni, corrispondenti alle combinazioni di agenzia-partito-elezione.

Un'altra variabile risposta

- Utilizzare un singolo punto della serie storica ignora l'**evoluzione temporale** delle intenzioni di voto, specie in prossimità delle elezioni.
- Un istituto che rileva un **trend** discendente o crescente nelle settimane precedenti all'elezione potrebbe essere penalizzato o favorito dal fatto che l'ultimo sondaggio disponibile non rifletta tale andamento.
- Per mitigare questi limiti si utilizza una stima interpolata della serie dei sondaggi fino al giorno dell'elezione, ottenuta tramite il **filtro di Kalman**.
- Questo approccio consente di incorporare in maniera coerente l'informazione contenuta nell'intera traiettoria dei sondaggi.

La serie storica

- Le risposte di ogni intervistato possono essere considerate prove di **Bernoulli indipendenti**.
- Aggregando queste risposte, un sondaggio sulle intenzioni di voto per un singolo partito alle prossime elezioni è una v.c. **Binomiale**.
- La serie storica per ogni agenzia e partito è suddivisa in tre sezioni, ognuna in corrispondenza di una elezione.
- **Esempio**. La prima sezione va dalle Politiche del 2018 alle Europee del 2019. Applicando il filtro si ottiene la stima delle intenzioni di voto per ogni partito e agenzia per le elezioni del 2019.
- Le ragioni di questa scelta sono due:
  1. I sondaggi pubblicati posteriormente ad un'elezione non servono per stimare l'errore legato quella stessa;
  2. Il valore della percentuale di voto di un determinato partito e agenzia per l'elezione all'inizio della sezione viene usato come **valore iniziale** per il filtro di Kalman;

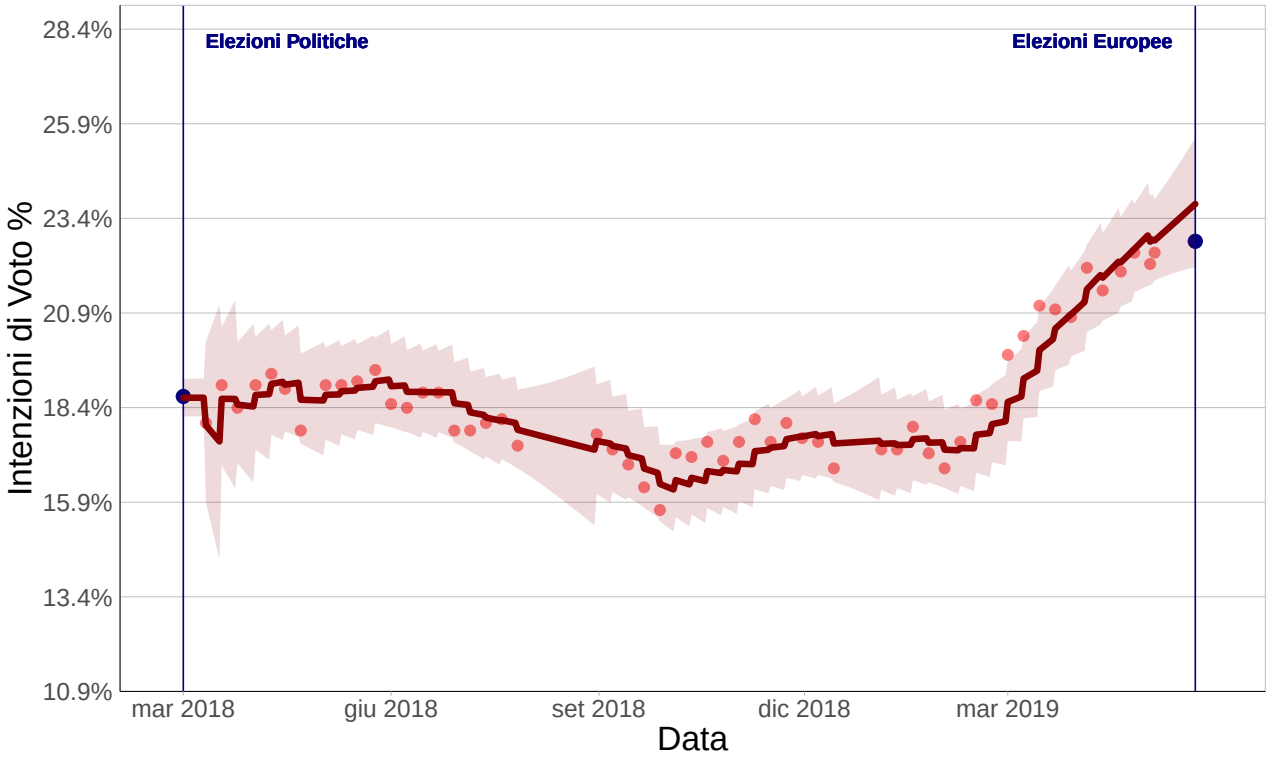
Modello state space Binomiale

- Si assume l'esistenza, per ogni giorno t, di un **vero valore** di intenzione di voto $\pi_t \in (0, 1)$ che evolve nel tempo in modo regolare.
- Si implementa un modello **state-space binomiale** con le intenzioni di voto che si evolvono secondo un **local linear trend**:

$$
\begin{aligned}
p(y_t \mid \theta_t) &= \text{Bin}\left(u_t, \pi_t = \frac{\exp\{\theta_t\}}{\exp\{1 + \theta_t\}}\right), \quad && u_t = \text{campione}_t \\
\theta_{t+1} &= \theta_t + \nu_t + \eta_t, && \eta_t \sim N(0, \sigma_\eta^2) \\
\nu_{t+1} &= \nu_t + \xi_t && \xi_t \sim N(0, \sigma_\xi^2)
\end{aligned}
$$

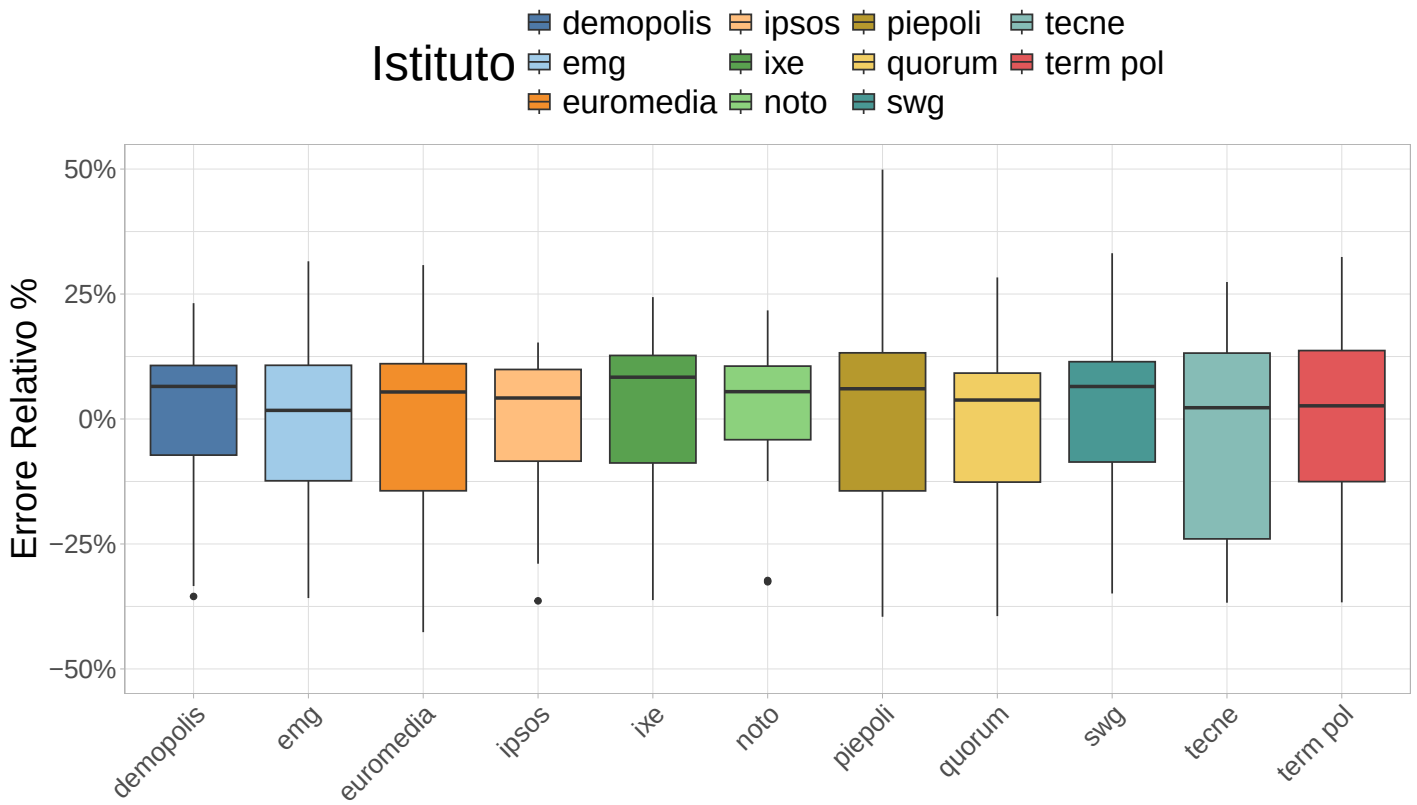
- Il **logit** delle intenzioni di voto di domani è uguale a quello di oggi più uno shock casuale e il termine ν_t che si evolve nel tempo come un **random walk**.
- La dinamica è flessibile e regolare tale da permettere di colmare razionalmente i giorni senza sondaggi, sfruttando l'informazione dei giorni precedenti.

Esempio: PD - Swg - Europee 2019



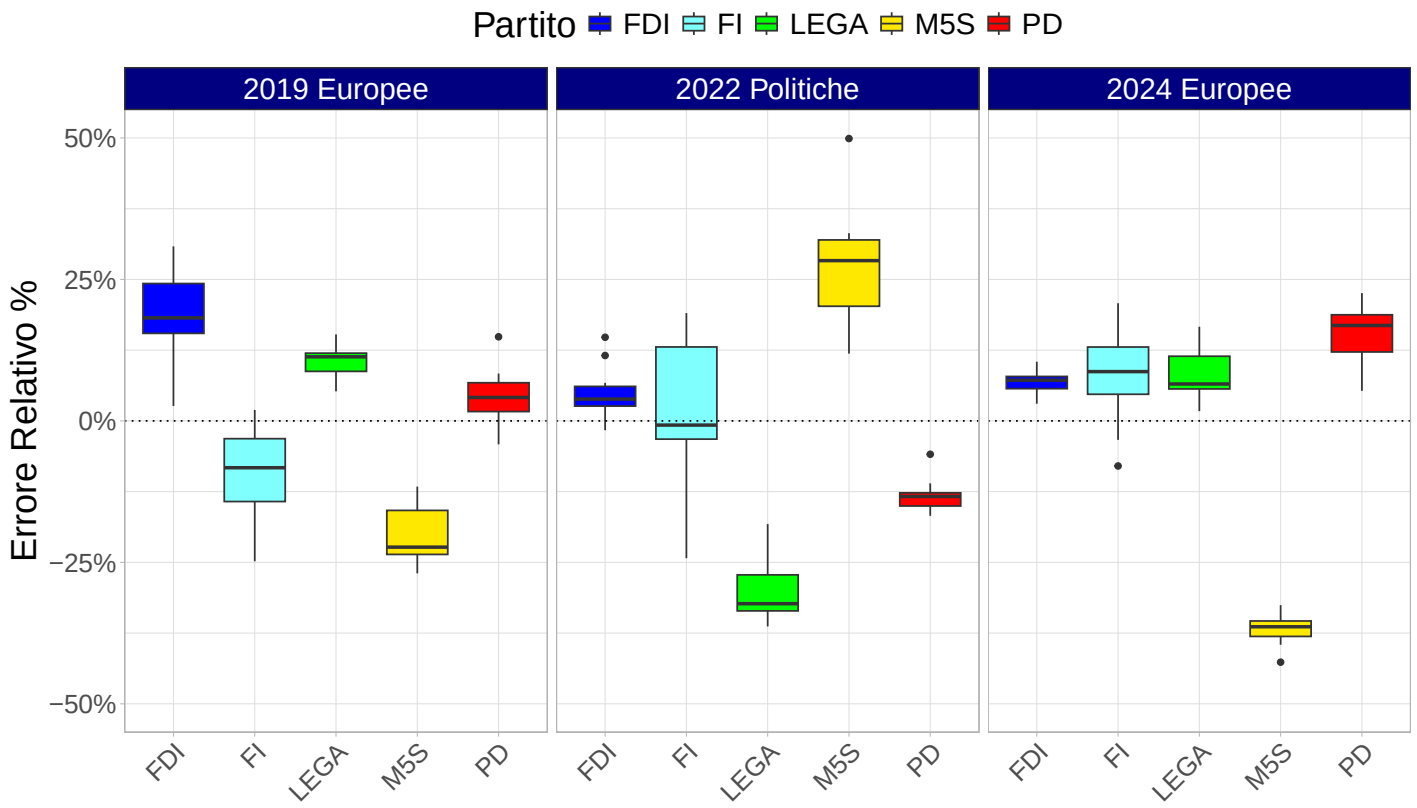
- Stima filtrata del livello delle intenzioni di voto per il partito PD nel primo intervallo d'interesse per l'agenzia Swg.
- Nei periodi senza osservazioni, il filtro si basa esclusivamente sulle **dinamiche predittive** del modello.
- Una volta che una nuova osservazione diventa disponibile, il filtro si **ricalibra**, correggendo eventuali deviazioni accumulate.

Boxplot variabile risposta per `Istituto`



- Boxplot della variabile risposta, definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, per `Istituto`.
- **Intuizione**. Non emergono **sistematicità** evidenti, suggerendo che le agenzie demoscopiche non differiscano metodicamente in termini di errore relativo.

Boxplot variabile risposta per `Partito` ed `Elezione`



- Boxplot della variabile risposta per ogni combinazione di `Partito` - `Elezione`.
- **Intuizione**. Compaiono chiari segnali di **sistematicità**, indicando che tali fattori contribuiscono a spiegare la variabilità dell'errore.

Modellistica e ANOVA I

- Si adotta un **modello lineare**:

$$\texttt{rel.err}_i = \beta_0 + \beta_1 \, \texttt{Istituto}_i + \beta_2 \, \texttt{Partito}_i + \beta_3 \, \texttt{Elezione}_i + \beta_4 \, (\texttt{Partito}_i \times \texttt{Elezione}_i).$$

  con $i = 1, \ldots, 165$.

- L'**ANOVA** consente di valutare l'importanza relativa delle diverse fonti di variazione presenti: i fattori `Istituto`, `Partito`, `Elezione` e l'interazione fra quest'ultimi.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Istituto	10	0.0418	0.00418	0.9234	0.5137
Partito	4	0.7314	0.18286	40.3507	$< 2 \cdot 10^{-16}$
Elezione	2	0.0211	0.01056	2.3306	0.1010
Partito:Elezione	8	4.2531	0.53163	117.3138	$< 2 \cdot 10^{-16}$
Residui	140	0.6344	0.00453		

- **Risultato chiave**. `Istituto` non fornisce un contributo informativo rilevante e la sua inclusione non migliora in modo significativo la capacità del modello di spiegare la variabile risposta.

- **Risultato chiave**. `Partito` e l'interazione hanno p-value estremamente piccolo: si rifiuta l'ipotesi nulla che sostiene che non abbiano effetto sulla risposta.

- Appare ragionevole stimare un nuovo modello lineare che escluda la variabile indipendente `Istituto`:

$$\begin{aligned}\texttt{rel.err}_i = \beta_0 &+ \beta_1 \texttt{Partito}_i \\ &+ \beta_2 \texttt{Elezione}_i + \beta_3 (\texttt{Partito}_i \times \texttt{Elezione}_i).\end{aligned}$$

con $i = 1, \ldots, 165$.

Eteroschedasticità degli errori

- **Problema**. Dall'analisi grafica dei residui emerge la possibile presenza di eteroschedasticità.
- **Problema.** Gli usuali test possono risultare inappropriati, dunque si potrebbe arrivare a conclusioni inferenziali errate.
- **Soluzione**. Si confrontano i modelli attuando un test di Wald, utilizzando una matrice di varianza – covarianza robusta all'eteroschedasticità.
- **Risultato chiave**. L'ipotesi nulla non viene rifiutata: la variabile `Istituto` non migliora in modo statisticamente significativo la capacità esplicativa del modello.

Inferenza sulle medie previste

- L'attenzione si sposta sugli **intervalli di confidenza** delle medie previste della risposta per ogni combinazione di `Partito` ed `Elezione`.
- **Obiettivo**. Verificare se tali medie risultino statisticamente diverse da zero.
- É necessario ricorrere ad uno **stimatore robusto** della matrice di covarianza:

$$\widehat{Var}(\hat{\beta}) = (X^\top X)^{-1} X^\top \text{diag}\left[\frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}\right] X (X^\top X)^{-1},$$

- Per ciascuna combinazione delle variabili esplicative `Partito` ed `Elezione`, identificata dal vettore c, si è interessati a testare l'ipotesi nulla

$$H_0 : c^\top \beta = 0.$$

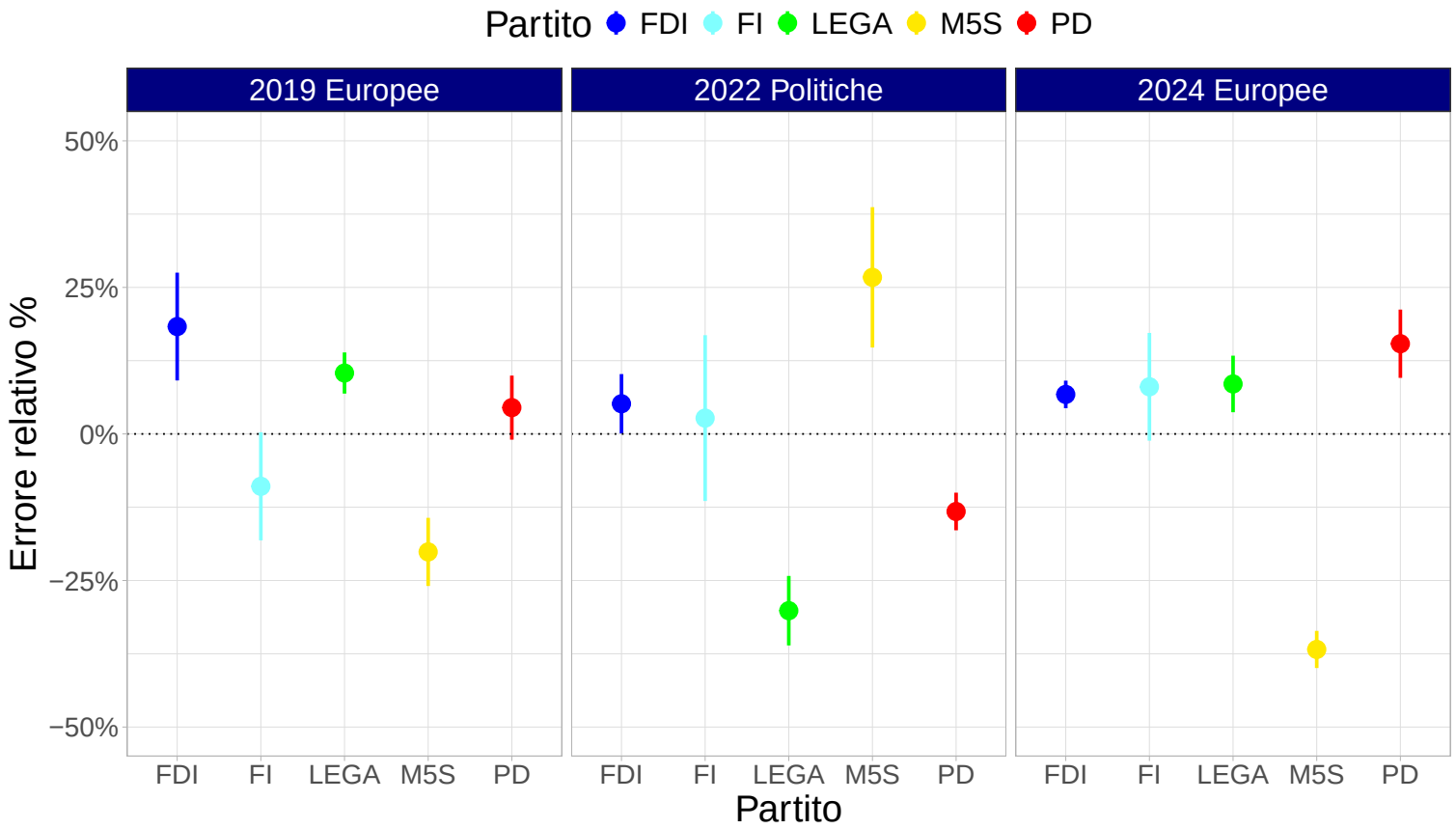
- Quindi si costruisce il seguente **intervallo di confidenza**:

$$\text{IC}_{1-\alpha}(c^\top \hat{\beta}) = c^\top \hat{\beta} \pm \; t_{1-\alpha/2,\, n-k} \sqrt{c^\top \widehat{\text{Var}}(\hat{\beta}) c}.$$

Problema delle comparazioni multiple

- **Problema**. Comparazioni multiple, la probabilità di osservare almeno un rifiuto erroneo dell'ipotesi nulla cresce all'aumentare del numero di ipotesi testate.
- Le comparazioni sono tante quante le combinazioni `Partito -Elezione`, cioè $5 \times 3 = 15$.
- Nel presente studio si è adottato l'approccio più semplice e diretto, ossia la correzione di Bonferroni.
- **Idea**. Il **livello di significatività** prefissato α viene diviso per il **numero di test** condotti m, ottenendo una soglia di significatività corretta pari a α/m.
- **Limite.** La correzione di Bonferroni può risultare piuttosto conservativa.
- La riduzione della probabilità di incorrere in errori di Tipo I comporta un aumento della probabilità di commettere errori di Tipo II.

Intervalli di confidenza corretti per Bonferroni

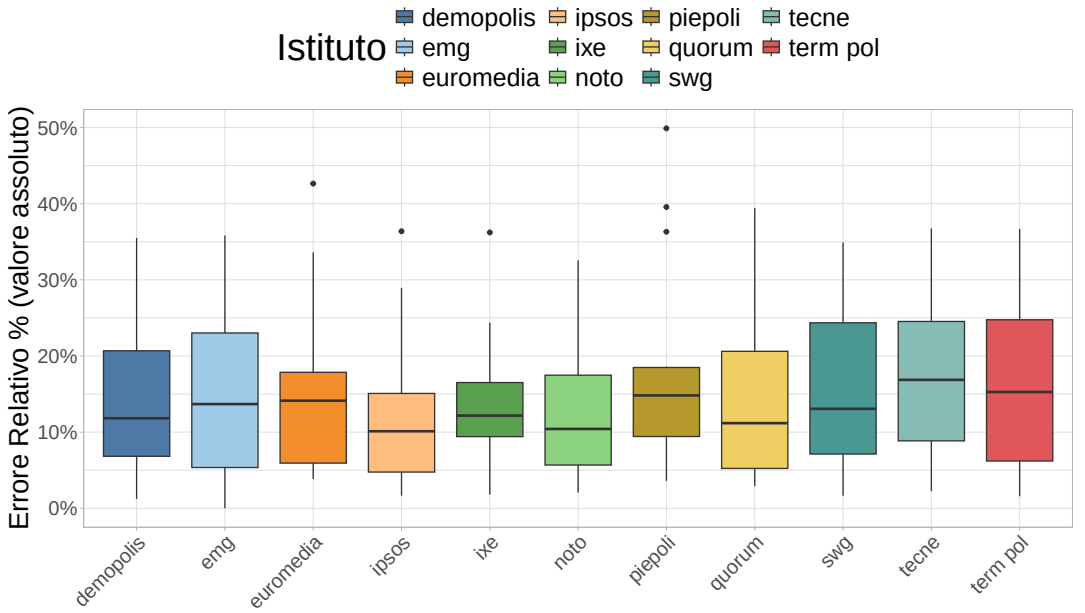


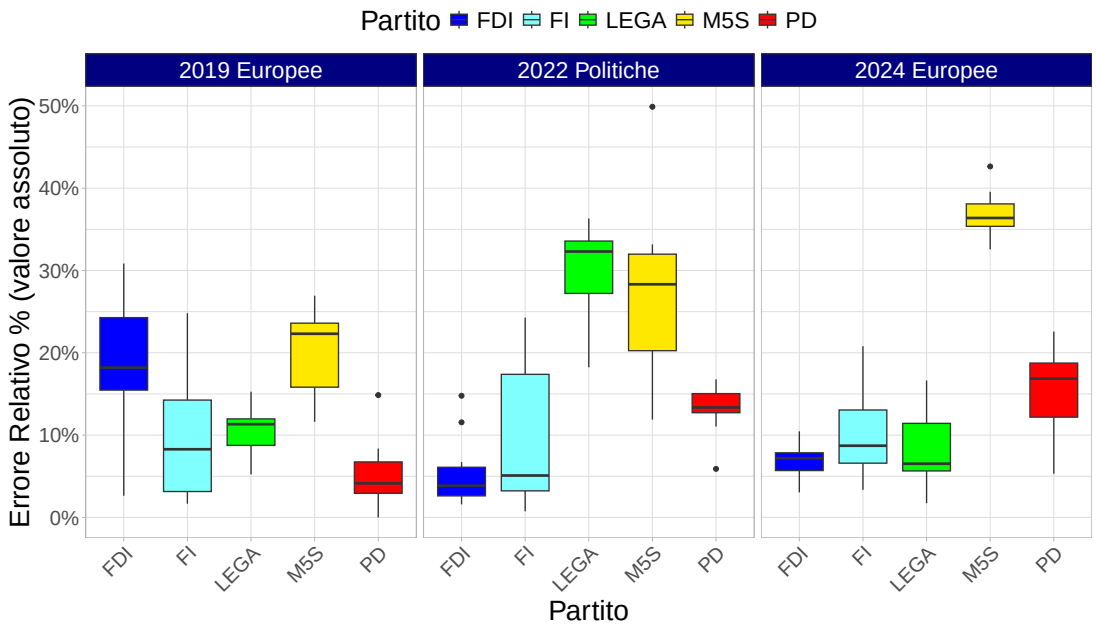
- **Risultato chiave**. Vi è evidenza della presenza di **distorsioni sistematiche** specifiche per partito ed elezione.
- Tali schemi riflettono caratteristiche strutturali della stima delle intenzioni di voto provenienti da sondaggi pre-elettorali.

Inferenza sull'eterogeneità delle agenzie

- Ora ci si concentra sulla **differenza di affidabilità** tra le agenzie demoscopiche.
- Precedentemente si distingueva tra errori di sovrastima ed errori di sottostima. Ora tale distinzione risulta ridondante e secondaria.
- **Obiettivo**. Valutare la precisione complessiva delle stime fornite dalle diverse agenzie.
- Si considera il **valore assoluto** dell'errore relativo: sovrastime e sottostime della stessa entità vengono trattate in modo equivalente.

Analisi descrittiva della variabile risposta





Modellistica

- Vengono stimati due modelli lineari classici:

$$\begin{aligned}\text{abs}(\texttt{rel.err}_i) = \beta_0 &+ \beta_1 \texttt{Istituto}_i + \beta_2 \texttt{Partito}_i \\ &+ \beta_3 \texttt{Elezione}_i + \beta_4 (\texttt{Partito}_i \times \texttt{Elezione}_i)\end{aligned}$$

$$\begin{aligned}\text{abs}(\texttt{rel.err}_i) = \beta_0 &+ \beta_1 \texttt{Partito}_i + \beta_2 \texttt{Elezione}_i \\ &+ \beta_3 (\texttt{Partito}_i \times \texttt{Elezione}_i)\end{aligned}$$

- Si vuole testare l'ipotesi che la variabile `Istituto` non fornisca un contributo informativo rilevante capace di spiegare la variabile risposta.
- **Problema**. L'analisi grafica dei residui suggerisce la possibile presenza di eteroschedasticità.
- **Soluzione**. É necessario adottare una procedura robusta all'eteroschedasticità, specularmente a quanto fatto precedentemente
- Dal test di Wald robusto si conclude che l'ipotesi nulla non viene rifiutata: la covariata `Istituto` non è statisticamente significativa.

Commento al risultato

- Una volta controllato per il contesto politico ed elettorale attraverso le variabili `Partito`, `Elezione` e la loro interazione, le diverse agenzie demoscopiche non mostrano differenze sistematiche in termini di affidabilità.

- L'errore commesso nelle stime delle intenzioni di voto non dipende dall'istituto che realizza il sondaggio, ma dal contesto specifico in cui la previsione viene formulata.

- **Risultato chiave**. Le agenzie demoscopiche selezionate per quest'analisi risultano sostanzialmente **intercambiabili** in termini di affidabilità: a parità di informazione e di contesto, un'agenzia vale l'altra.

Discussione dei risultati

- **Sintesi**. I partiti vengono stimati per ogni elezione con un certo errore che non è specifico del singolo istituto, ma condiviso.
- Questo risultato consente di formulare una critica all'approccio del Poll of Polls.
- L'**aggregazione** dei sondaggi non elimina questa componente di **distorsione condivisa**, che rimane inalterata anche quando si media su un gran numero di osservazioni.
- Un modello aggregativo è della stessa qualità dei sondaggi di cui è composto: se tutte le agenzie sono sistematicamente in errore nella stessa direzione, il processo di aggregazione produrrà risultati fuorvianti.
- Il Poll of Polls non rappresenta una soluzione ai limiti strutturali dei sondaggi pre-elettorali.
- In Shirani-Mehr et al. (2018) si dimostra che, nel contesto americano, i sondaggi relativi a una stessa elezione tendono a condividere una distorsione sistematica comune.