Explore the linear relationship between positive rate and the coverage rate of voters, effectiveness of vote for a poll

Assignment 2

Jianheng Chen - 1005680746

2021.10.22

Introduction

Abstract

In the modern society, poll is one of the most popular methods to collect data or people's preference in different contexts. The scale of polls can vary from "whether to have a dinner in the newly opened restaurant" by a mom to her kids or "who will you choose to be your next president" for all citizens. For example, by researching the 2004 US presidential election's poll it indicates that both newspaper-based and web-based polls, even entirely differ in data collection process, were playing important roles in following the votes will in a comprehensive system(Ben-Ur & Newman, 2010). Some popular news/broadcasting companies will even hold poll checkers for the important events. According to CBC News (2021), there are more than 20 websites, such as forum research and abacus data, are creating polls to predict the winner of Canada's 2021 election, showing an overall trend of Liberal will win the most seats but not the majority while Justin Trudeau is likely to win the election.

However, polls are not always correct and sometimes can be misleading or confusing for several reasons. Beyond most of the prediction institutions' expectation, Trump won the 2016 US election with a comparative explicitly lower support rate than Clinton. By studying the poll conducted by ThirtyFiveEight and Upshot, surprisingly, Wright. F and Wright. A(2018) found that these websites ignored the low coverage rate of the polls. It means that the ballots are not fully distributed to all of the valid voters. By researching in detail Wright. F and Wright. A found high level educated white voters are likely to vote for Democrat but in the mean time a large number of potential voters including non-college educated white and people from other ethic groups are not covered in the poll(not receive the ballots or excluded from the poll even they have citizenship) that their opinions are excluded(2018). The coverage rate of voters can significantly bias the result of the poll.

Moreover, the response rate of the poll, which is the number of ballots returned to the pollsters with respect to the ballots they distribute has great impact on pool's quality and persuasiveness. After studying 36 surveys and polls provided by several pollsters from 2007 to 2009, Farag and Aly conclude that the result of the survey will be badly biased by loss of response(2009). Furthermore, they concluded that low response rate can even affect the correctness of the model that the data they collected were not valid to follow the true willingness of the voters.

These two problems will be discussed related to the opinions of the voters in the dataset provided in the following part.

Dataset

The data I collected is from City Toronto(2021) about polls for whether to construct public utilities conducted by Toronto city government. The information collected are id of poll application, address, type of application for public utilities, number of ballots distributed, number of ballots returned, number of ballots showing in favor or oppose, the data opening date, the number of potential voters and whether the response rate met the requirement indicating the poll's validness.

I choose this dataset as it is generally believed that residents are always concerned with public utilities construction. For the people showing favor in the polls they might think new utilities bring them some conveniences in life. However, on the other hand the people oppose in the votes may have considerations in safety, noise and environment problem brought from the construction.

The Toronto Transit Commission, known as TTC, is a company operating and monitoring the transportation system in Toronto. Interestingly, TTC has been criticized for the extension for its line 1. As there was growing need and votes for subway line in the northern Toronto area such as Vaughan and Sheppard west, the Toronto City Council approved the extension permit for TTC line 1 in 2005(Wikipedia, 2021). However, even a poll showed that the residents there approve the construction of new subway line the voice of oppose of stores and parking lot holders were ignored. By December 2017 the construction area was stilled consider to be surrounded by mainly big-box stores and parking lot(Wikipedia, 2021). The requirement for projected population was not fulfilled and a lot of criticism emerged as the stores and parking lot holders' business were badly affected. Indeed, Wikipedia concluded that the cost of operation became incredibly high as $30 million in 2016(2021).

By looking into this example, analysis for the quality of polls about constructions looks extremely necessary for checking the necessity of constructing new public utility and cost saving. In particular, after discussing the importance of coverage rate of potential voters and response rate for a poll, the information I aim to examine are the number of ballots showing in favor, the number of ballots returned, the the number of potential voters and the response rate met. To conclude, my research question will be **Is there a linear relationship between the positive rate of a poll and its coverage rate as well as effectiveness of the poll?**.

To be specific, the positive rate of the poll is the **number of "in favor" ballots/the number of ballots returned**, the coverage rate will be **number of ballots distributed/amount of potential voters** and the effectiveness of the poll is whether the ballots returned meet the response rate is distinguish by **yes** or **no** in categories.

Hypothesis

For the hypothesis part, I assume the coverage rate will have a negative relationship with the positive rate of polls. As the remaining potential voters are considered to be independent with the people already distributed with ballots, they may not have the same trend of voting as the people already covered. By saying this, increasing the coverage rate will decrease the positive rate of polls.

What is more, the higher the response rate met I assume the higher positive rate of the poll will be. The response rate met basically measure if the voters are enthusiastic enough to vote for the construction of public utility. If the response rate is not enough the poll will be counted as not valid that if the voters really want the government to construct something there is a trend for them to use the ballots instead of giving no response.

Guideline

This report will discuss the problem in several parts. In the **Data** section how the data is collected and some useful variables in the research question will be introduced. The **Method** part will be focusing on why do I choose such a model and the assumptions will be discussed. The execution of the model and what I get from the model will be in the **Result** section and a **conclusion** will be made at the end of this report.

Data

Data Collection Process

The dataset is created by the City Clerk's Office and the data collection was started from April 1,2015, which most recently get undated on October 20, 2021(City Toronto, 2021). The data is titled with "Polls regarding changes in a neighbourhood" discussing the opinions of property owners, residents and businesses owner for newly applications for public utilities construction. To be specific, the constructions include Boulevard Cafe, Off-street parking, permit parking, traffic calming and business improvement area(City Toronto, 2021). According to the City(2021), when an application is submitted the City will use three ways of polling to different types of citizens in the specific area, which are ePolling, Information Regarding Boulevard Cafe, Parking & Traffic Polls and Information Regarding BIA Polls.

ePolling An email or text message will be sent to all types of citizens in the area affected by the public construction(City of Toronto, 2017).

Information Regarding Boulevard Cafe, Parking & Traffic Polls The resident or tenant of a property over 18 years old will be mailed a notice giving information about the poll, a postage-paid return envelope and a ballot with a deadline on it and the result of the poll will be available after 10 business of the deadline(City of Toronto, 2017).

Information Regarding BIA Polls The business improvement area(BIA) are areas that business property owner, commercial and industrial tenant promoting economic development(City of Toronto, 2020). These three types of people will receive the BIA ballot.

After all these types of polls were finished a final approval will be decided by the City Council and the dataset will be created to store the results.

Drawback

One foreseeable drawback of this data collecting method is the difficulty in identifying coincide polls. Even creating three ways of collecting data will benefit from covering more potential voters in the area, it is hard to identify if someone is distributed with multiple kinds of polls and submit repeatedly. It significantly increase the workload of data cleaning(to check if the voters ID appear multiple times).

Data Cleaning

As describe in the **Dataset** part the information collected are id of poll application, address, type of application for public utilities, number of ballots distributed, number of ballots returned, number of ballots showing in favor or oppose, the data opening date, the number of potential voters and whether the response rate was met.

The next step we will do is to clean the data. Firstly, the dataset has a several of columns indicating the id of applications for construction, which is irrelevant to the research question I want to discuss, so they will firstly be removed. Similarly, the declaration, the poll result, the pass rate label, the addresses of application, the information about date of the polls and the application types will be consider as residuals as they did not help to find the result of the question. Moreover, as we want to find the positive rate of the polls, the number of ballots showing "in favor" and the number of ballots returned to the pollsters will remain in the dataset while the number of ballots "oppose" or need further processing will be removed.

To clarify, there is a column called ballots returned to sender but it in actual means the ballots that did not find the voters' address and get mailed back and the column named received by voters means the number of ballots distributed minus the number of ballots did not find the voters. We do not really need this information so they will also be removed. Next, there are too many blanks in the column of pass rate and its algorithm is unclear that it will be deleted as well.

Finally, we check the polls information by rows to identify if there are any outliers or blanks. Fortunately there is not too many blanks(only 9) so we can remove them easily. Moreover, all of the remaining columns will be renamed with shorter names, which is not a must but brings a lot of conveniences. After this, we will create new variables of positive rate and coverage rate using the formulas in **Dataset** part.

Important variables

After cleaning the data and creating the new variables we need in the research question, there are overall 7 variables in the new dataset. Reminded that the research question is the linear relationship between the positive rate of a poll and its coverage rate as well as effectiveness of the polls, there are three main variables of our interest and concerned.

Positive rate of the poll This variable measures the proportion of voters agree or show favor in the construction of new utility to all the voters giving valid feedback.

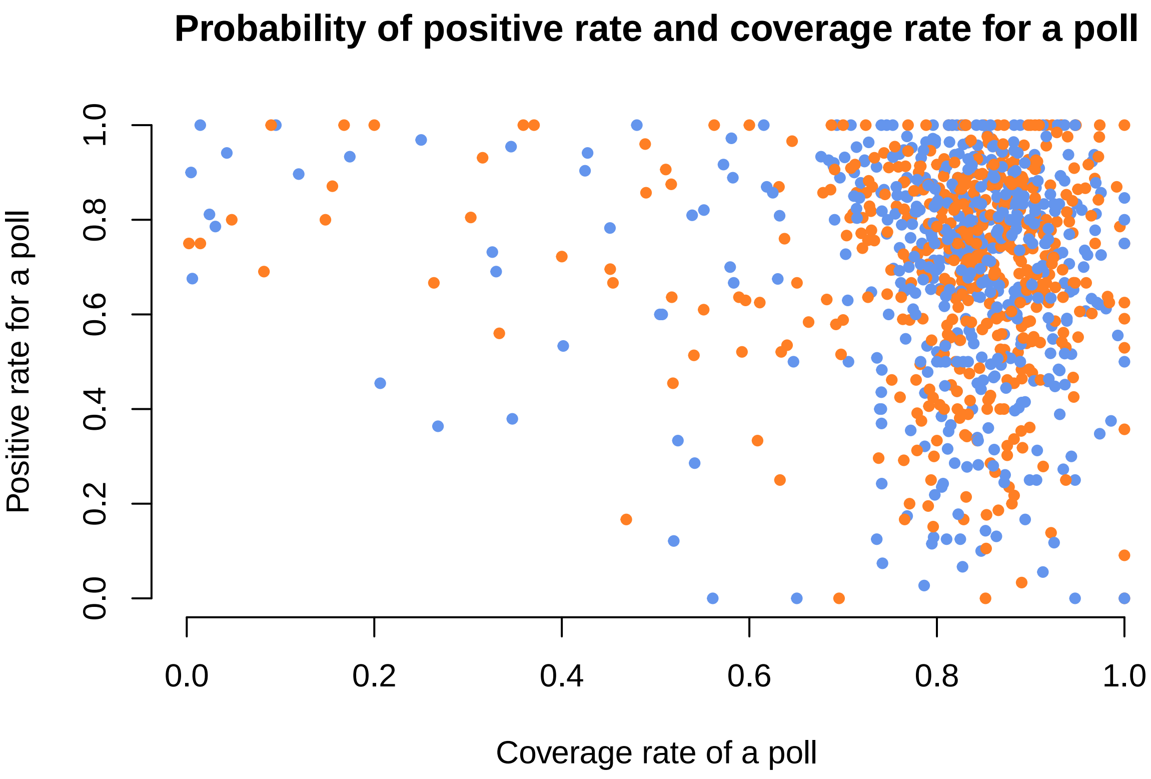
coverage rate The variable measures how well the ballots capture the opinions of all potential voters in specific areas.

requirement met This is a variable describing if the number of ballots return met the criteria to prove the validness of the result of the poll.

Table 1: Distribution of positive rate of polls and the coverage rate of polls

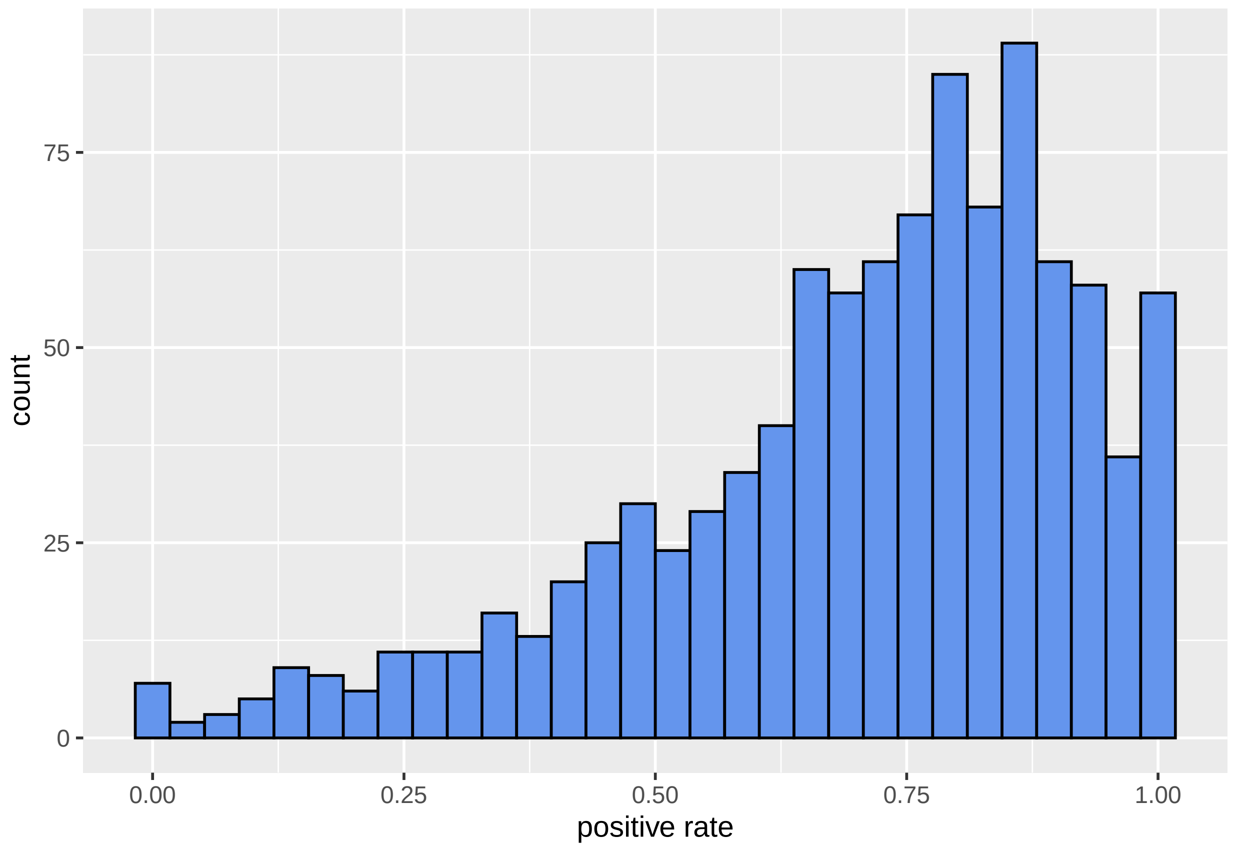
	vars	n	mean	sd	min	max	range	se
positive_rate	1	1003	0.7045726	0.2176721	0.0000000	1	1.0000000	0.0068731
coverage_rate	2	1003	0.8108324	0.1478256	0.0023183	1	0.9976817	0.0046677

Graph 1: Probability of positive rate and coverage rate for a poll



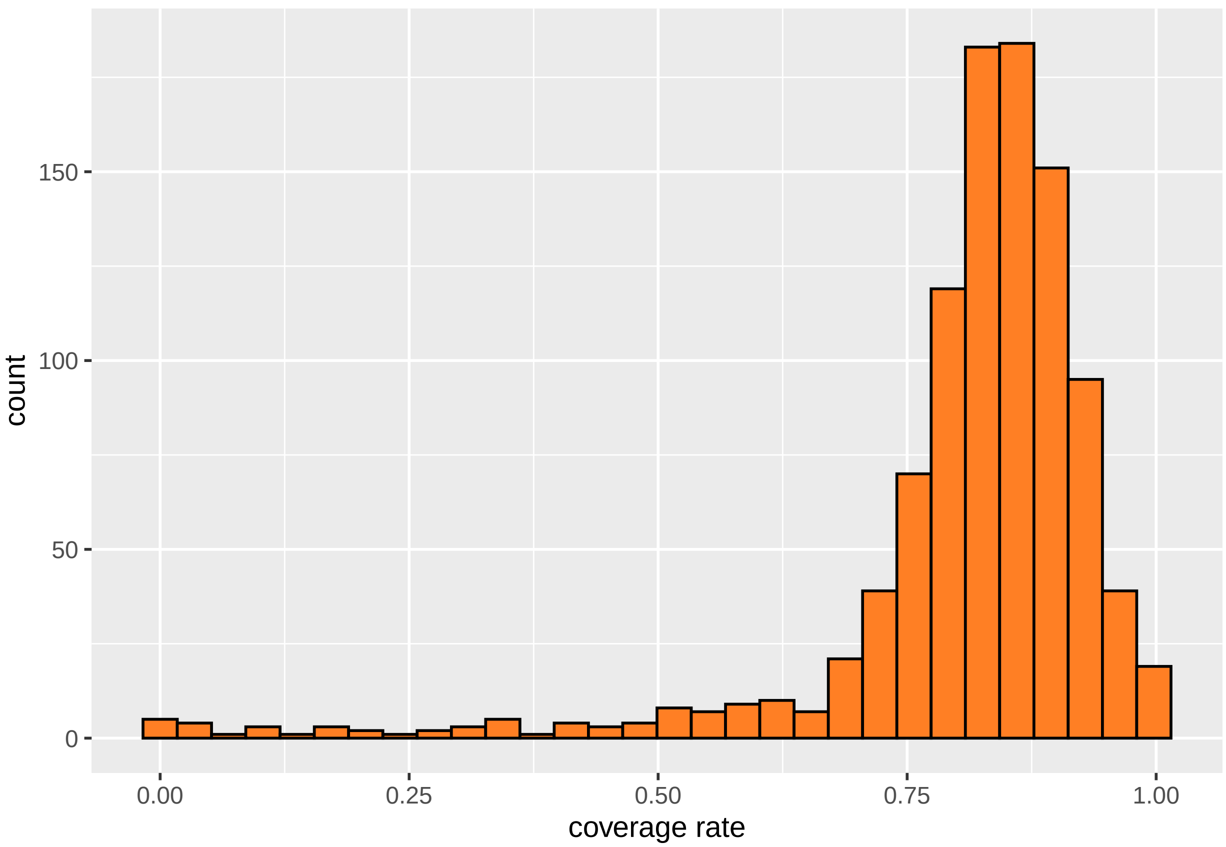
As shown in the above results(blue points stand for coverage rate and chocolate point stand for positive rate in the graph), based on the 1003 observations we find that the average positive rate of 0.705 and coverage rate of 0.811 are comparatively high(as we can see the points concentrated at the right top).

Graph 2: Distribution of positive rates for all polls



With this average rate and a standard deviation of 0.217(showed in **table 1**) means most of the voters are approving the construction of new utilities. The result also showed that the standard deviation of coverage rate is 0.148 which means the distribution is pretty left skewed as showed in the histogram.

Graph 3: Distribution of coverage rates for all polls



According to the **graph 3**, the City Council has done a great job in covering most of the potential voters. However, the minimum coverage rate is 0.23% as showed in **table 1** for an area and the significantly left skewedness means the ballots distribution still can be refined.

As for the requirement met, which is a categorical variable to split the observations to a binary groups(Wikipedia, 2021) For the probability graph of it please see **Appendix**.

All analysis for this report was programmed using `R version 4.1.1` with `tidyverse 1.3.1` package (Wickham et al., 2021) and `pysch 2.1.9` package

Methods

Methodology

For the construction of model we first need to figure out which methodology is the most appropriate for our model. Noted that our research question aims to find the relationship between positive rate of a poll and coverage rate, whether the required response was met we need to choose between linear model and logistic model. According to Mondal(2020), the linear regression focus on the numerically relationship of dependent variable and independent variables while logistic regression is to clarify elements to a set into two groups. For our case, the dependent variable we want to study is numerical so choosing linear regression model is suitable to get the answer of the question.

A further classification will be made for whether to choose from Bayesian model or Frequentist model. The frequentist model assume the data is sampled from some distribution and the parameters of interest where the Bayesian model assume the parameter is also a variable and follow some distribution(Causevic, 2020). Indeed, to measure how the change in 1 unit of coverage rate or valid polls with respect to positive rate of a poll there is no necessity to consider the parameter of them as variables given the cleaned dataset. In conclusion, a frequentist linear regression model will be built.

Methematical model

Here we will construct a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \epsilon$$

. This model can directly describe the linear relationship based on the parameters of intercept and slope. Based on the **Introduction** section we care about how can we improve in making the agreement on constructing new utilities more reliable. The dependent variable, y is the positive rate of polls where the two independent variables x_1, x_2 stands for coverage rate of polls and validness of the polls. Note that the coverage rate is a numerical variable and validness of polls is a categorical variable as discussed in the **important variables**.

Next, the fist parameter β_0 here represents the intercept of the regression line that the value of positive rate when the coverage rate is 0 and the response rate no met given residuals. The parameter β_1 is measuring how the positive rate will change with respect to change in one unit of coverage rate. What is more, as the validness of poll is categorical, β_2 plays the role of what changes will be there if the poll is valid(printed as **yes** in the variable "requirement met").

Moreover, what we are going to do is to choose an appropriate model selection technique. The first thing we need to keep in our mind is that we want to check if a linear relationship exist between the dependent variable and independent variables. To be specific, our null hypothesis is $H_0 : \beta_1 = 0$ and our alternative hypothesis is $H_A : \beta_1 \neq 0$ and our goal is to reject the null hypothesis and the same process will run for β_2 as well. Rumsey claimed in her article(2021) that if we want to test whether to reject a hypothesis for a population(which means all the observations here), p-value is a good choice. What a p-value do here is to measure if our result is significant enough to reject the null hypothesis, which is suitable for our model. What is more, we will choose a 5% level of confidence, which means that if the p-value is smaller than 0.05 we are sufficient enough to reject H_0.

Results

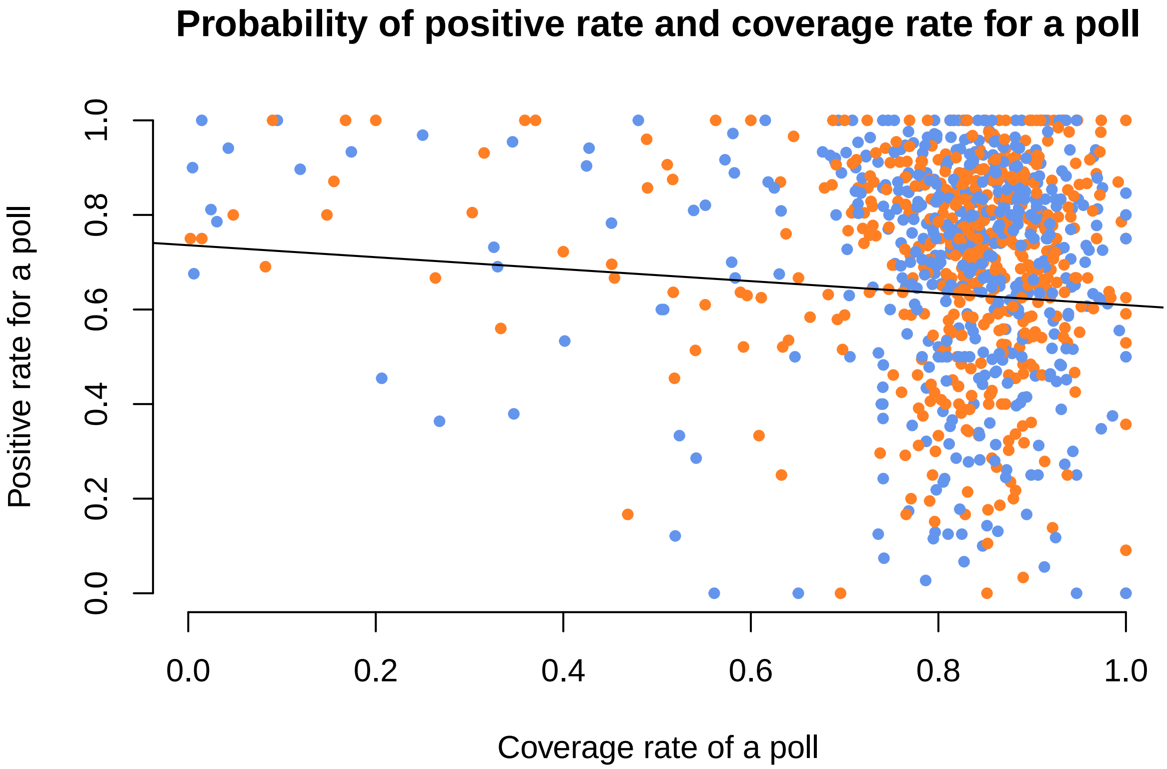
Table 2: Results of the linear regression model

Parameter	*Estimate*	*Std.Error*	$t - value$	Pr>$\|t\|$
$\hat{\beta}_0$	0.736	0.040	18.254	0.000
$\hat{\beta}_1$	-0.127	0.046	-2.761	0.059
$\hat{\beta}_2$	0.087	0.018	4.956	0.000

To our first recognition, the intercept is 0.736. Remind that in the **Method** part we discussed the parameter's character according to the two independent variables, coverage rate of polls and requirement met. From table 2 we can clearly see that if the coverage rate increase by 1% there will be 0.127% drop in the expected positive rate of poll within the variation of 0.046%, which is proving our hypothesis in the **Introduction** section. As if the poll meet the requirement, the expected positive rate will go up by 0.087 by average within a scale of 0.018, which is also coincide with our assumption.

Recalled that we choose a 5% confidence level in **Method**, it is time to verify if we need to reject the null hypothesis. Again, for $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$ we find the p-value correspond to the null hypothesis is 0.059, which is higher than the confidence level 0.05. With this result we can not say there is evidence to show positive rate of polls has significant relationship with coverage rate. Next, for $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$, the p-value is nearly 0 with 3 decimal places, which means we can claim there is a significant relationship for positive rate and validness of polls.

Graph 4: Probability of positive rate and coverage rate for a poll with linear model



The the result of response met is quite reasonable as if we reconsider the Graph 2, the distribution of positive rate was quite concentrated around 0.75, if the poll is valid the expected positive rate of poll is tend to increase as the validness of poll strengthen the trend. With the support of t-value we can believe there is a significant relationship between positive rate and validness of a poll. Noted that it will only change the intercept but the slope of the regression is based on coverage rate. According to Graph 4, the slope of the regression is not apparent(in our result we know it is -0.127). It is hardly we can say a change in percentage point of coverage rate will greatly impact the result of positive rate(especially when both of them are fractional number), which conform to what we got in the comparison of p-value and 5% confidence level.

All analysis for this report was programmed using `R version 4.1.1`. I used the `lm()` function in base `R` to derive the estimates of a frquentist linear regression in this section.

Conclusions

Reveled the importance of polls in modern society, we came up with a research question **Is there a linear relationship between the positive rate of a poll and its coverage rate as well as effectiveness of the poll?**. In order to solve it, we built a frequentist linear model and use p-value to identify if we have significant relationships between dependent variable and independent variables. As far as the research goes we conclude that there is a significant relationship in positive rate of polls and validness of a poll. However, by comparing the p-value of the parameter of coverage rate with the confidence level of interest, we failed to find a clear relevance between it and positive rate of polls.

Weaknesses

For to the process of this research, the biggest obstacle I met is that there are too many observations need to deal with. Researcher must be extremely careful on the data cleaning process that the research can move on. What is more, there is a weakness about consideration for graph making. Due to the number of observations it is hardly to make a perfect look table/graph we can clearly figure out the linear relationship that a model selection technique should also be well considered.

Next Steps

As far as the research goes I recognize the importance of a categorical variable in the original dataset, the types of applications of polls. In the **result** section there is no significant relationship between the positive rate of polls and the coverage rate of polls, but what will happen if we can distribute them into different groups by types of applications. It is possible that we will find the linear relationship can exist in some categories of applications.

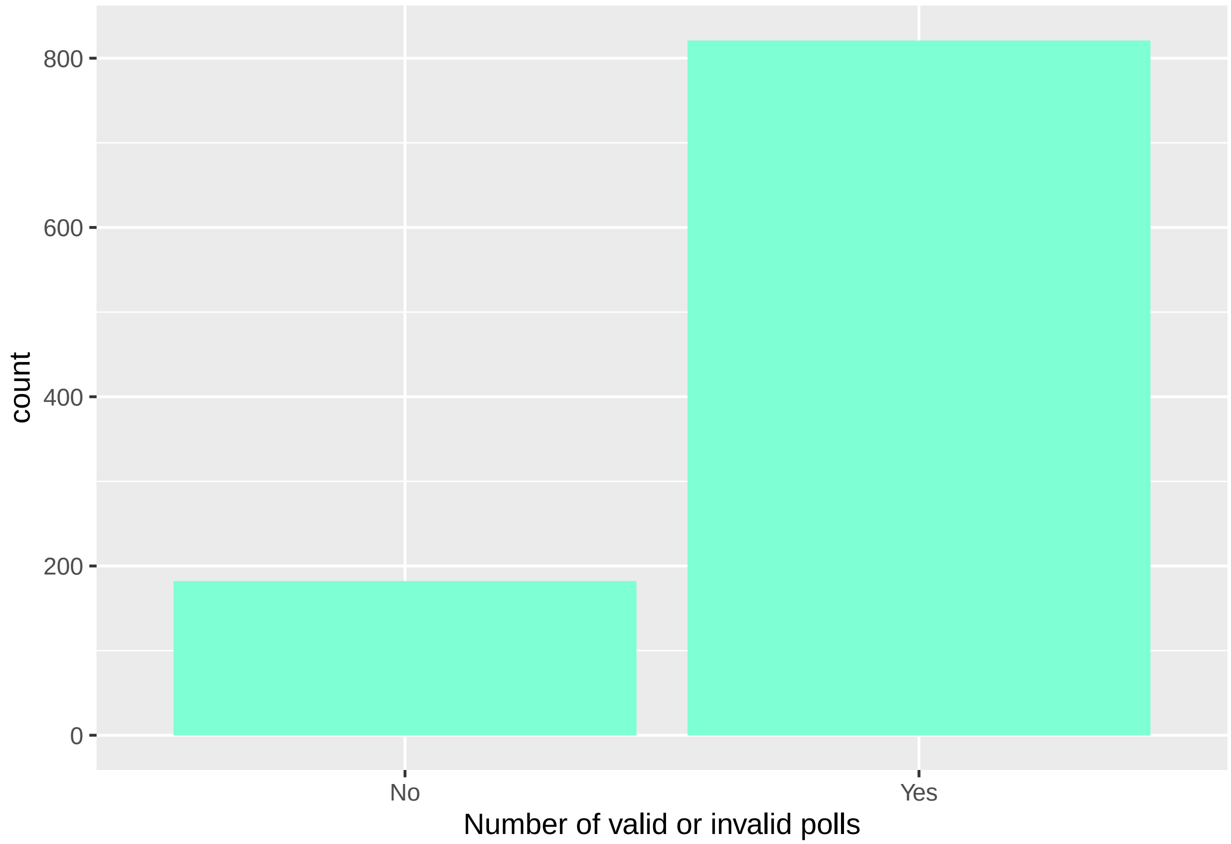
Discussion

By studying Ben-Ur(2010), Wright, F and Wright A(2018), Farag and Aly(2009)'s research we know the importance of coverage rate of polls and whether the requirement rate is met for a poll. We created a regression to determine the significance of their relationships with positive rate of poll when there are still more factors we can have attempt on.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)
2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)
4. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*
5. City Toronto. (2015, April 1). *Open data dataset.* City of Toronto Open Data Portal. Retrieved October 16, 2021, from https://open.toronto.ca/dataset/polls-conducted-by-the-city/.
6. CBC News. (2021, September 19). *CBC News Canada poll tracker.* CBCnews. Retrieved October 20, 2021, from https://newsinteractives.cbc.ca/elections/poll-tracker/canada/.
7. Ben-Ur, J., & Newman, B. I. (2010). A marketing poll: an innovative approach to prediction, explanation and strategy. *European Journal of Marketing.*
8. Wright, F. A., & Wright, A. A. (2018). How surprising was Trump's victory? Evaluations of the 2016 US presidential election and a new poll aggregation model. *Electoral Studies*, 54, 81-89.
9. Farag, A. R. A., & Aly, H. M. (2009). Response Rate at the Public Opinion Poll Center.
10. Wikipedia. (2021, October 18). *Line 1 yonge–university. Wikipedia.* Retrieved October 20, 2021, from https://en.wikipedia.org/wiki/Line_1_Yonge%E2%80%93University.
11. City of Toronto. (2017, December 1). *Polls regarding changes in a neighbourhood.* City of Toronto. Retrieved October 21, 2021, from https://www.toronto.ca/city-government/planning-development/polls-regarding-changes-in-a-neighbourhood/.
12. City of Toronto. (2020, September 22). Business Improvement Areas (bias). City of Toronto. Retrieved October 21, 2021, from https://www.toronto.ca/city-government/accountability-operations-customer-service/city-administration/city-managers-office/agencies-corporations/agencies/business-improvement-areas-bias/.
13. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4 (43), 1686.
14. Revelle, W. (2021). *psych: Procedures for Psychological*, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.1.9, https://CRAN.R-project.org/package=psych
15. Wikipedia. (2021, September 17). *Categorical variable.* Wikipedia. Retrieved October 21, 2021, from https://en.wikipedia.org/wiki/Categorical_variable.
16. Mondal, S. (2020, December 2). *Linear vs logistic regression: Linear and logistic regression.* Analytics Vidhya. Retrieved October 21, 2021, from https://www.analyticsvidhya.com/blog/2020/12/beginners-take-how-logistic-regression-is-related-to-linear-regression/.
17. Causevic, S. (2020, August 30). *Frequentist vs. bayesian approaches in Machine Learning.* Medium. Retrieved October 21, 2021, from https://towardsdatascience.com/frequentist-vs-bayesian-approaches-in-machine-learning-86ece21e820e.
18. Rumsey, D. J. (2021, July 13). *How to determine a P-value when testing a null hypothesis.* dummies. Retrieved October 21, 2021, from https://www.dummies.com/education/math/statistics/how-to-determine-a-p-value-when-testing-a-null-hypothesis/.

Appendix


800
600
400
200
0
count
No
Yes
Number of valid or invalid polls