Forecasting the 2024 U.S. Presidential Election*

My subtitle if needed

Tim Chen Steven Li Tommy Fu

October 19, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

```
# Load necessary libraries
library(dplyr)
library(tidyr)

# Load the dataset
data <- read.csv(here::here("data/02-analysis_data/analysis_data.csv"))
```

Simple Linear Regression for Kamala Harris and Donald Trump

Kamala Harris We will perform a simple linear regression using pct (poll percentage) as the dependent variable and sample_size as the independent variable.

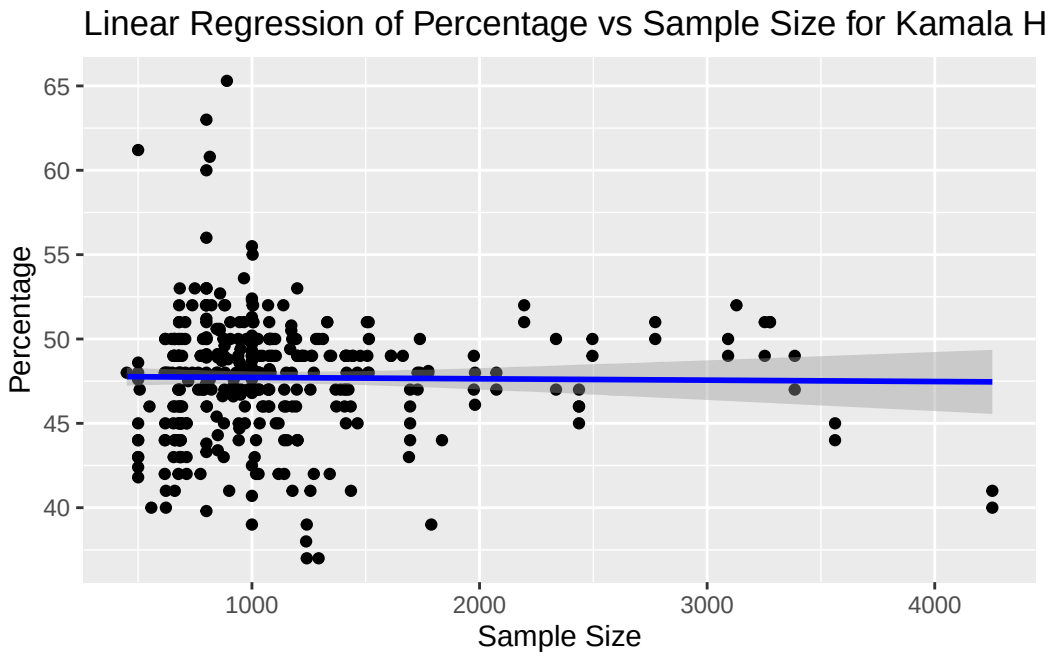
```
# Load necessary library
library(ggplot2)

# Filter the data for only "Kamala Harris"
harris_data <- subset(data, candidate_name == "Kamala Harris")

# Perform a simple linear regression with pct as the dependent variable and sample_size as th
model <- lm(pct ~ sample_size, data = harris_data)

# Summary of the model
summary(model)
```

```
Call:
lm(formula = pct ~ sample_size, data = harris_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7014  -1.7436   0.2814   2.2516  17.5702

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.780e+01  3.819e-01 125.171   <2e-16 ***
sample_size -8.092e-05  3.025e-04  -0.268    0.789
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.53 on 388 degrees of freedom
Multiple R-squared:  0.0001844, Adjusted R-squared:  -0.002392
F-statistic: 0.07157 on 1 and 388 DF,  p-value: 0.7892
```

```r
# Plot the relationship for Kamala Harris
ggplot(harris_data, aes(x = sample_size, y = pct)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression of Percentage vs Sample Size for Kamala Harris",
       x = "Sample Size",
       y = "Percentage")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Donald Trump We will repeat the same process for Donald Trump.

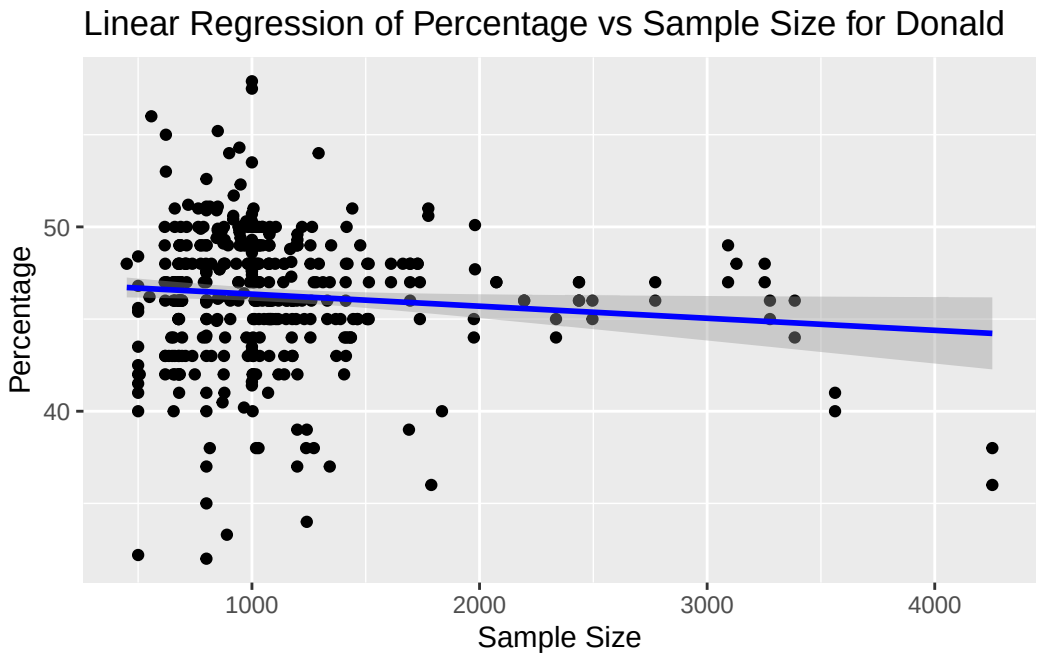
```r
# Load necessary library
library(ggplot2)

# Filter the data for only "Kamala Harris"
```

```r
trump_data <- subset(data, candidate_name == "Donald Trump")

# Perform a simple linear regression with pct as the dependent variable and sample_size as t
model <- lm(pct ~ sample_size, data = trump_data)

# Summary of the model
summary(model)
```

```
Call:
lm(formula = pct ~ sample_size, data = trump_data)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4840  -1.5635   0.4427   2.4530  11.5470

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.0077785  0.3904934 120.380   <2e-16 ***
sample_size -0.0006547  0.0003114  -2.102   0.0361 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.637 on 409 degrees of freedom
Multiple R-squared:  0.01069,   Adjusted R-squared:  0.008273
F-statistic:  4.42 on 1 and 409 DF,  p-value: 0.03612
```

```r
# Plot the relationship for Kamala Harris
ggplot(trump_data, aes(x = sample_size, y = pct)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Linear Regression of Percentage vs Sample Size for Donald",
       x = "Sample Size",
       y = "Percentage")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Multiple Linear Regression (MLR) for Kamala Harris and Donald Trump

Kamala Harris In this step, we will build a multiple linear regression (MLR) model for Kamala Harris using additional predictors like numeric_grade, pollscore, transparency_score, and state.

```
# Convert categorical variables to factors for Kamala Harris
harris_data$state <- as.factor(harris_data$state)
harris_data$methodology <- as.factor(harris_data$methodology)

# Build MLR model for Kamala Harris
mlr_harris_model <- lm(pct ~ numeric_grade + pollscore + transparency_score + sample_size + s

# Summary of the MLR model for Harris
summary(mlr_harris_model)
```

```
Call:
lm(formula = pct ~ numeric_grade + pollscore + transparency_score +
    sample_size + state + methodology, data = harris_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7299 -1.2580  0.2084  1.5162  6.4101
```

Coefficients:

	Estimate
(Intercept)	43.7364837
numeric_grade	1.6229500
pollscore	1.3887700
transparency_score	-0.0308776
sample_size	0.0003072
stateCalifornia	12.1754527
stateConnecticut	5.4029969
stateFlorida	-4.0137925
stateGeorgia	0.4347585
stateIndiana	-7.9813842
stateIowa	-2.3304007
stateMaryland	16.6524107
stateMassachusetts	14.4329701
stateMichigan	1.1401156
stateMinnesota	3.3670869
stateMissouri	-5.0692548
stateMontana	-7.9313842
stateNational	0.8488022
stateNebraska	-8.0126599
stateNebraska CD-2	4.7454955
stateNevada	1.3076078
stateNew Hampshire	1.3686158
stateNew Mexico	5.1210480
stateNew York	5.8025425
stateNorth Carolina	1.5080407
stateOhio	-1.7541512
statePennsylvania	1.7334175
stateRhode Island	6.9029969
stateTexas	-1.8251108
stateVirginia	3.5260505
stateWisconsin	2.6039604
methodologyLive Phone	-1.2537774
methodologyLive Phone/Email	-0.9965708
methodologyLive Phone/Online Panel/Text	0.7400844
methodologyLive Phone/Online Panel/Text-to-Web	0.9902808
methodologyLive Phone/Text-to-Web	0.3976375
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	-0.0235439
methodologyOnline Ad	0.2245155
methodologyOnline Panel	-1.1730826
methodologyOnline Panel/Text-to-Web	1.6749094

methodologyProbability Panel	-1.2376664
	Std. Error
(Intercept)	7.2553925
numeric_grade	2.6906121
pollscore	0.7860164
transparency_score	0.3263822
sample_size	0.0002802
stateCalifornia	2.7165827
stateConnecticut	2.6780552
stateFlorida	1.2504132
stateGeorgia	0.6913505
stateIndiana	2.7172373
stateIowa	2.6475167
stateMaryland	2.7167289
stateMassachusetts	1.6538015
stateMichigan	0.7009881
stateMinnesota	1.3175397
stateMissouri	1.9156162
stateMontana	2.0481444
stateNational	0.6501052
stateNebraska	1.9613148
stateNebraska CD-2	1.2021938
stateNevada	0.9140826
stateNew Hampshire	2.0481444
stateNew Mexico	1.5849699
stateNew York	1.3799780
stateNorth Carolina	0.6740308
stateOhio	1.2580954
statePennsylvania	0.6455982
stateRhode Island	1.9958699
stateTexas	1.1910107
stateVirginia	1.1865896
stateWisconsin	0.6841719
methodologyLive Phone	0.7973025
methodologyLive Phone/Email	1.4665033
methodologyLive Phone/Online Panel/Text	2.5896119
methodologyLive Phone/Online Panel/Text-to-Web	0.8530883
methodologyLive Phone/Text-to-Web	0.8643934
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	2.1597049
methodologyOnline Ad	0.9238667
methodologyOnline Panel	0.8486324
methodologyOnline Panel/Text-to-Web	1.0073697
methodologyProbability Panel	1.0892340

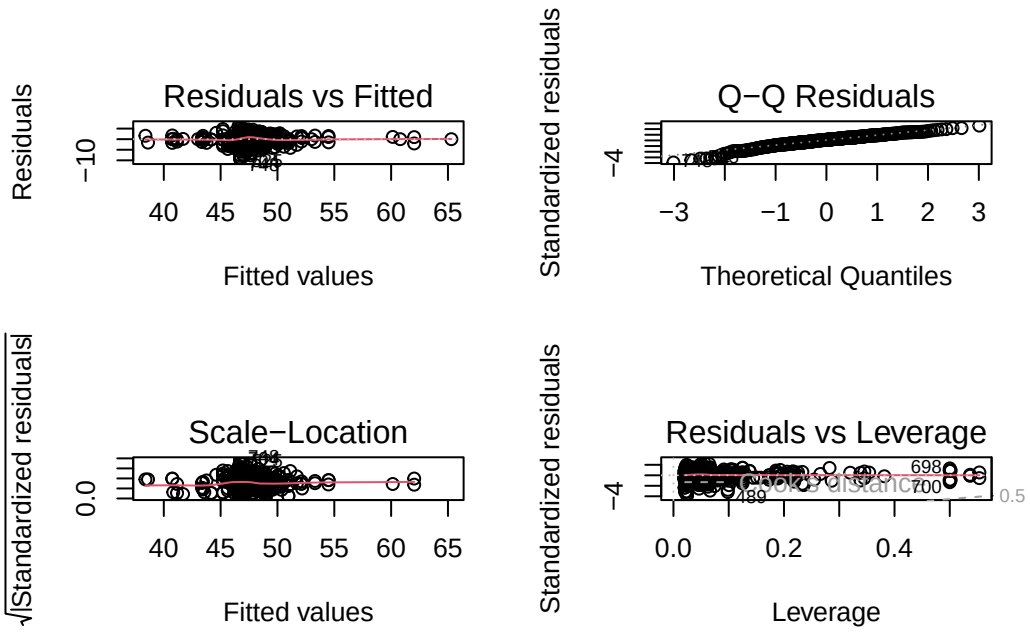
	t value
(Intercept)	6.028
numeric_grade	0.603
pollscore	1.767
transparency_score	-0.095
sample_size	1.097
stateCalifornia	4.482
stateConnecticut	2.018
stateFlorida	-3.210
stateGeorgia	0.629
stateIndiana	-2.937
stateIowa	-0.880
stateMaryland	6.130
stateMassachusetts	8.727
stateMichigan	1.626
stateMinnesota	2.556
stateMissouri	-2.646
stateMontana	-3.872
stateNational	1.306
stateNebraska	-4.085
stateNebraska CD-2	3.947
stateNevada	1.431
stateNew Hampshire	0.668
stateNew Mexico	3.231
stateNew York	4.205
stateNorth Carolina	2.237
stateOhio	-1.394
statePennsylvania	2.685
stateRhode Island	3.459
stateTexas	-1.532
stateVirginia	2.972
stateWisconsin	3.806
methodologyLive Phone	-1.573
methodologyLive Phone/Email	-0.680
methodologyLive Phone/Online Panel/Text	0.286
methodologyLive Phone/Online Panel/Text-to-Web	1.161
methodologyLive Phone/Text-to-Web	0.460
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	-0.011
methodologyOnline Ad	0.243
methodologyOnline Panel	-1.382
methodologyOnline Panel/Text-to-Web	1.663
methodologyProbability Panel	-1.136
	Pr(>|t|)

```
(Intercept)                                                     4.22e-09 ***
numeric_grade                                                   0.546774
pollscore                                                       0.078128 .
transparency_score                                              0.924682
sample_size                                                     0.273556
stateCalifornia                                                 1.00e-05 ***
stateConnecticut                                                0.044408 *
stateFlorida                                                    0.001451 **
stateGeorgia                                                    0.529856
stateIndiana                                                    0.003531 **
stateIowa                                                       0.379345
stateMaryland                                                   2.38e-09 ***
stateMassachusetts                                               < 2e-16 ***
stateMichigan                                                   0.104758
stateMinnesota                                                  0.011024 *
stateMissouri                                                   0.008507 **
stateMontana                                                    0.000129 ***
stateNational                                                   0.192536
stateNebraska                                                   5.46e-05 ***
stateNebraska CD-2                                              9.56e-05 ***
stateNevada                                                     0.153464
stateNew Hampshire                                              0.504433
stateNew Mexico                                                 0.001351 **
stateNew York                                                   3.32e-05 ***
stateNorth Carolina                                             0.025894 *
stateOhio                                                       0.164117
statePennsylvania                                               0.007600 **
stateRhode Island                                               0.000610 ***
stateTexas                                                      0.126328
stateVirginia                                                   0.003168 **
stateWisconsin                                                  0.000167 ***
methodologyLive Phone                                           0.116735
methodologyLive Phone/Email                                     0.497236
methodologyLive Phone/Online Panel/Text                         0.775209
methodologyLive Phone/Online Panel/Text-to-Web                  0.246509
methodologyLive Phone/Text-to-Web                               0.645789
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone 0.991308
methodologyOnline Ad                                            0.808135
methodologyOnline Panel                                         0.167757
methodologyOnline Panel/Text-to-Web                             0.097279 .
methodologyProbability Panel                                    0.256622
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 2.525 on 349 degrees of freedom
Multiple R-squared:  0.5397,    Adjusted R-squared:  0.4869
F-statistic: 10.23 on 40 and 349 DF,  p-value: < 2.2e-16
```

```r
# Diagnostic plots for the MLR model
par(mfrow = c(2, 2))
plot(mlr_harris_model)
```

```
Warning: not plotting observations with leverage one:
  149, 180, 182, 195, 214, 218
```




```r
par(mfrow = c(1, 1))  # Reset plot layout
```

Similarly, we build an MLR model for Donald Trump.

```r
# Convert categorical variables to factors for Trump
trump_data$state <- as.factor(trump_data$state)
trump_data$methodology <- as.factor(trump_data$methodology)

# Build MLR model for Donald Trump
mlr_trump_model <- lm(pct ~ numeric_grade + pollscore + transparency_score + sample_size + st
```

```r
# Summary of the MLR model for Trump
summary(mlr_trump_model)
```

```
Call:
lm(formula = pct ~ numeric_grade + pollscore + transparency_score +
    sample_size + state + methodology, data = trump_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.3352 -1.2111  0.2025  1.4630  5.6933

Coefficients:
                                                                Estimate
(Intercept)                                                   5.251e+01
numeric_grade                                                -2.053e+00
pollscore                                                     1.036e+00
transparency_score                                            5.741e-01
sample_size                                                   1.725e-04
stateCalifornia                                              -1.248e+01
stateConnecticut                                             -1.088e+01
stateFlorida                                                  2.588e+00
stateGeorgia                                                  1.275e-01
stateIndiana                                                  7.392e+00
stateIowa                                                    -5.041e-02
stateMaryland                                                -1.719e+01
stateMassachusetts                                           -1.462e+01
stateMichigan                                                -1.525e+00
stateMinnesota                                               -4.213e+00
stateMissouri                                                 5.911e+00
stateMontana                                                  4.992e+00
stateNational                                                -1.860e+00
stateNebraska                                                 4.830e+00
stateNebraska CD-2                                           -5.856e+00
stateNevada                                                  -9.399e-01
stateNew Hampshire                                           -6.008e+00
stateNew Mexico                                              -5.695e+00
stateNew York                                                -8.420e+00
stateNorth Carolina                                          -6.978e-01
stateOhio                                                     2.004e+00
statePennsylvania                                            -1.466e+00
stateRhode Island                                            -6.379e+00
```

stateTexas	1.826e+00
stateVirginia	-6.486e+00
stateWisconsin	-1.055e+00
methodologyLive Phone	-2.890e+00
methodologyLive Phone/Email	-3.915e+00
methodologyLive Phone/Online Panel/Text	-2.954e-02
methodologyLive Phone/Online Panel/Text-to-Web	-1.291e+00
methodologyLive Phone/Text-to-Web	-2.213e+00
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	-3.752e+00
methodologyOnline Ad	4.338e-01
methodologyOnline Panel	-3.641e+00
methodologyOnline Panel/Text-to-Web	8.974e-01
methodologyProbability Panel	-6.305e+00
	Std. Error
(Intercept)	6.082e+00
numeric_grade	2.225e+00
pollscore	6.940e-01
transparency_score	2.990e-01
sample_size	2.571e-04
stateCalifornia	2.525e+00
stateConnecticut	2.488e+00
stateFlorida	1.162e+00
stateGeorgia	6.425e-01
stateIndiana	2.526e+00
stateIowa	2.457e+00
stateMaryland	2.525e+00
stateMassachusetts	1.536e+00
stateMichigan	6.375e-01
stateMinnesota	1.050e+00
stateMissouri	1.780e+00
stateMontana	1.904e+00
stateNational	5.964e-01
stateNebraska	1.815e+00
stateNebraska CD-2	1.117e+00
stateNevada	8.496e-01
stateNew Hampshire	1.904e+00
stateNew Mexico	1.470e+00
stateNew York	1.280e+00
stateNorth Carolina	6.262e-01
stateOhio	1.169e+00
statePennsylvania	5.938e-01
stateRhode Island	1.853e+00
stateTexas	1.106e+00

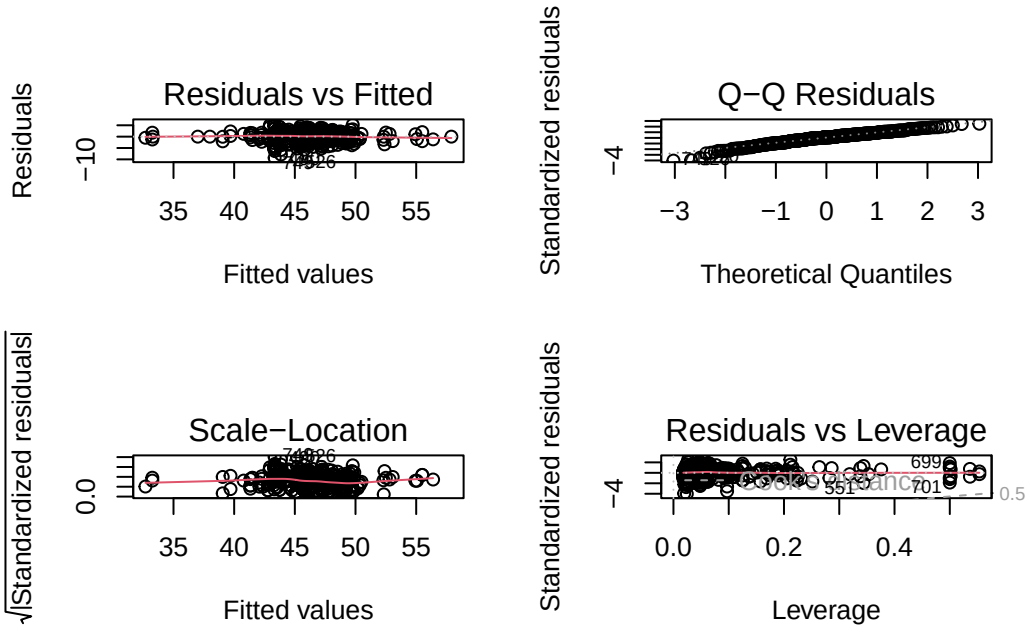
stateVirginia	1.099e+00
stateWisconsin	6.238e-01
methodologyLive Phone	7.364e-01
methodologyLive Phone/Email	1.353e+00
methodologyLive Phone/Online Panel/Text	2.406e+00
methodologyLive Phone/Online Panel/Text-to-Web	7.910e-01
methodologyLive Phone/Text-to-Web	7.907e-01
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	2.001e+00
methodologyOnline Ad	8.476e-01
methodologyOnline Panel	7.783e-01
methodologyOnline Panel/Text-to-Web	9.356e-01
methodologyProbability Panel	9.938e-01
	t value
(Intercept)	8.634
numeric_grade	-0.923
pollscore	1.493
transparency_score	1.920
sample_size	0.671
stateCalifornia	-4.941
stateConnecticut	-4.373
stateFlorida	2.227
stateGeorgia	0.198
stateIndiana	2.927
stateIowa	-0.021
stateMaryland	-6.807
stateMassachusetts	-9.521
stateMichigan	-2.391
stateMinnesota	-4.013
stateMissouri	3.321
stateMontana	2.622
stateNational	-3.118
stateNebraska	2.661
stateNebraska CD-2	-5.243
stateNevada	-1.106
stateNew Hampshire	-3.156
stateNew Mexico	-3.873
stateNew York	-6.578
stateNorth Carolina	-1.114
stateOhio	1.714
statePennsylvania	-2.469
stateRhode Island	-3.442
stateTexas	1.651
stateVirginia	-5.902

```
stateWisconsin                                                  -1.691
methodologyLive Phone                                           -3.924
methodologyLive Phone/Email                                     -2.895
methodologyLive Phone/Online Panel/Text                         -0.012
methodologyLive Phone/Online Panel/Text-to-Web                  -1.632
methodologyLive Phone/Text-to-Web                               -2.799
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone -1.875
methodologyOnline Ad                                             0.512
methodologyOnline Panel                                         -4.679
methodologyOnline Panel/Text-to-Web                              0.959
methodologyProbability Panel                                    -6.344
                                                               Pr(>|t|)
(Intercept)                                                     < 2e-16 ***
numeric_grade                                                  0.356805
pollscore                                                      0.136350
transparency_score                                             0.055616 .
sample_size                                                    0.502647
stateCalifornia                                                1.18e-06 ***
stateConnecticut                                               1.60e-05 ***
stateFlorida                                                   0.026557 *
stateGeorgia                                                   0.842825
stateIndiana                                                   0.003638 **
stateIowa                                                      0.983644
stateMaryland                                                  4.04e-11 ***
stateMassachusetts                                             < 2e-16 ***
stateMichigan                                                  0.017290 *
stateMinnesota                                                 7.27e-05 ***
stateMissouri                                                  0.000985 ***
stateMontana                                                   0.009098 **
stateNational                                                  0.001965 **
stateNebraska                                                  0.008137 **
stateNebraska CD-2                                             2.67e-07 ***
stateNevada                                                    0.269284
stateNew Hampshire                                             0.001731 **
stateNew Mexico                                                0.000127 ***
stateNew York                                                  1.63e-10 ***
stateNorth Carolina                                            0.265838
stateOhio                                                      0.087344 .
statePennsylvania                                              0.013992 *
stateRhode Island                                              0.000644 ***
stateTexas                                                     0.099572 .
stateVirginia                                                  8.13e-09 ***
stateWisconsin                                                 0.091662 .
```

```
methodologyLive Phone                                                  0.000104 ***
methodologyLive Phone/Email                                            0.004022 **
methodologyLive Phone/Online Panel/Text                                0.990207
methodologyLive Phone/Online Panel/Text-to-Web                         0.103618
methodologyLive Phone/Text-to-Web                                      0.005393 **
methodologyLive Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone      0.061638 .
methodologyOnline Ad                                                   0.609125
methodologyOnline Panel                                                4.05e-06 ***
methodologyOnline Panel/Text-to-Web                                    0.338137
methodologyProbability Panel                                           6.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.347 on 370 degrees of freedom
Multiple R-squared:  0.6271,    Adjusted R-squared:  0.5868
F-statistic: 15.56 on 40 and 370 DF,  p-value: < 2.2e-16
```

```r
# Diagnostic plots for the MLR model
par(mfrow = c(2, 2))
plot(mlr_trump_model)
```

```
Warning: not plotting observations with leverage one:
  149, 180, 182, 195, 214, 218
```



```r
par(mfrow = c(1, 1))  # Reset plot layout
```

Check Multicollinearity using Variance Inflation Factor (VIF)

```r
# Load the car package for VIF
if (!require(car)) install.packages("car")
```

```
Loading required package: car
```

```
Loading required package: carData
```

```
Attaching package: 'car'
```

```
The following object is masked from 'package:rstanarm':

    logit
```

```
The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```r
library(car)

# Check VIF for Kamala Harris model
vif(mlr_harris_model)
```

```
                      GVIF Df GVIF^(1/(2*Df))
numeric_grade     4.516076  1        2.125106
pollscore         3.049542  1        1.746294
transparency_score 7.163704  1        2.676510
sample_size       1.676196  1        1.294680
state            18.682289 26        1.057915
methodology      83.067087 10        1.247302
```

```r
# Refine the model by removing less significant predictors (e.g., methodology and transparenc
mlr_harris_model_refined <- lm(pct ~ numeric_grade + pollscore + sample_size + state, data =

# Summary of the refined Harris model
summary(mlr_harris_model_refined)
```

```
Call:
lm(formula = pct ~ numeric_grade + pollscore + sample_size +
    state, data = harris_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.1566  -1.3082   0.2477   1.6769   6.5569

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         51.2936767  5.3634319   9.564  < 2e-16 ***
numeric_grade       -1.2503473  2.0338873  -0.615 0.539103
pollscore            1.2204973  0.7107597   1.717 0.086808 .
sample_size          0.0003932  0.0002812   1.398 0.162926
stateCalifornia     14.1544541  2.6497822   5.342 1.64e-07 ***
stateConnecticut     5.8691676  2.6549788   2.211 0.027690 *
stateFlorida        -4.0333745  1.2721655  -3.170 0.001652 **
stateGeorgia         0.1755537  0.7095090   0.247 0.804717
stateIndiana        -6.0182800  2.6502047  -2.271 0.023745 *
stateIowa           -3.5860188  2.6577074  -1.349 0.178092
stateMaryland       18.6249673  2.6498304   7.029 1.05e-11 ***
stateMassachusetts  14.4315281  1.5926544   9.061  < 2e-16 ***
stateMichigan        0.7298417  0.7125376   1.024 0.306388
stateMinnesota       2.7755646  1.2877943   2.155 0.031802 *
stateMissouri       -4.4066179  1.9139854  -2.302 0.021887 *
stateMontana        -5.9682800  1.9090680  -3.126 0.001914 **
stateNational       -0.0255640  0.6097162  -0.042 0.966580
stateNebraska       -8.0699250  1.9259855  -4.190 3.51e-05 ***
stateNebraska CD-2   4.2204645  1.1805969   3.575 0.000398 ***
stateNevada          1.0323190  0.9354564   1.104 0.270527
stateNew Hampshire   3.3317200  1.9090680   1.745 0.081802 .
stateNew Mexico      4.9138463  1.5859557   3.098 0.002099 **
stateNew York        5.6569780  1.2868206   4.396 1.45e-05 ***
stateNorth Carolina  1.2991614  0.6870664   1.891 0.059443 .
stateOhio           -1.9849217  1.2769921  -1.554 0.120974
```

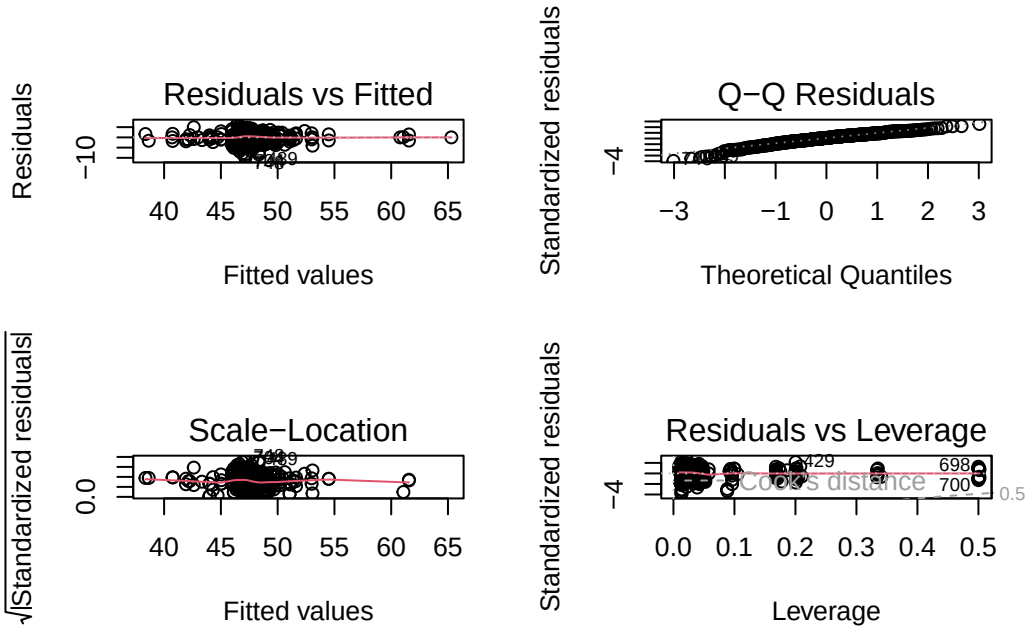
```
statePennsylvania    1.5149064  0.6549919   2.313 0.021293 *
stateRhode Island    7.3691676  1.9156899   3.847 0.000142 ***
stateTexas          -1.6514417  1.1844218  -1.394 0.164086
stateVirginia        3.7193287  1.1936069   3.116 0.001980 **
stateWisconsin       2.2092392  0.6669599   3.312 0.001019 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.6 on 360 degrees of freedom
Multiple R-squared:  0.4968,    Adjusted R-squared:  0.4563
F-statistic: 12.26 on 29 and 360 DF,  p-value: < 2.2e-16
```

```r
# Check VIF for the refined model
vif(mlr_harris_model_refined)
```

```
                 GVIF Df GVIF^(1/(2*Df))
numeric_grade 2.435023  1        1.560456
pollscore     2.352921  1        1.533923
sample_size   1.593471  1        1.262328
state         2.075947 26        1.014146
```

```r
# Plot diagnostic plots for the refined model
par(mfrow = c(2, 2))
plot(mlr_harris_model_refined)
```

```
Warning: not plotting observations with leverage one:
  149, 180, 182, 195, 214
```




```
par(mfrow = c(1, 1))  # Reset plot layout
```

Stepwise Model Selection for Optimization

```
# Perform stepwise selection to optimize the Harris model
step_model <- step(mlr_harris_model_refined, direction = "both")
```

```
Start:  AIC=773.97
pct ~ numeric_grade + pollscore + sample_size + state

                Df Sum of Sq    RSS    AIC
- numeric_grade  1      2.55 2435.5 772.38
<none>                      2432.9 773.97
- sample_size    1     13.21 2446.1 774.08
- pollscore      1     19.93 2452.8 775.15
- state         26   2171.97 4604.9 970.80

Step:  AIC=772.38
pct ~ pollscore + sample_size + state

                Df Sum of Sq    RSS    AIC
<none>                      2435.5 772.38
- sample_size    1     14.10 2449.6 772.63
```

```
+ numeric_grade  1      2.55 2432.9 773.97
- pollscore      1     65.63 2501.1 780.75
- state         26   2188.00 4623.5 970.37
```

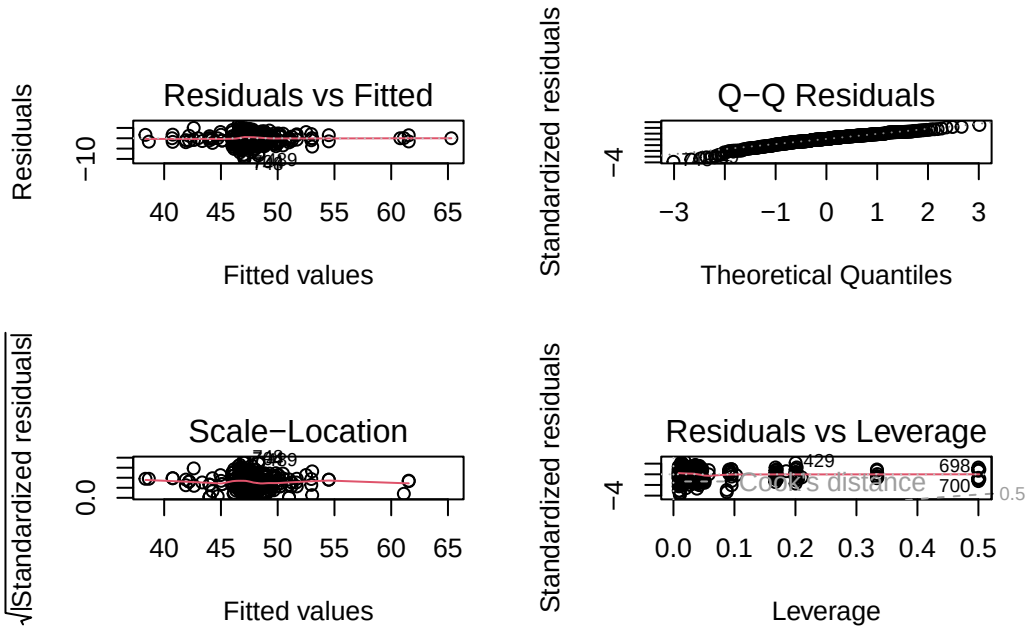
```r
# Summary of the stepwise model
summary(step_model)
```

```
Call:
lm(formula = pct ~ pollscore + sample_size + state, data = harris_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0749  -1.2596   0.1846   1.6866   6.4582

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         48.0354742  0.8218512  58.448  < 2e-16 ***
pollscore            1.5354101  0.4922962   3.119 0.001961 **
sample_size          0.0004052  0.0002803   1.446 0.149128
stateCalifornia     14.1232461  2.6470123   5.336 1.69e-07 ***
stateConnecticut     5.8687009  2.6526902   2.212 0.027567 *
stateFlorida        -4.0752071  1.2692493  -3.211 0.001443 **
stateGeorgia         0.1622771  0.7085689   0.229 0.818982
stateIndiana        -6.0517143  2.6473627  -2.286 0.022837 *
stateIowa           -3.4587875  2.6473530  -1.307 0.192212
stateMaryland       18.5928568  2.6470319   7.024 1.08e-11 ***
stateMassachusetts  14.4115807  1.5909513   9.058  < 2e-16 ***
stateMichigan        0.6886309  0.7087657   0.972 0.331904
stateMinnesota       2.9071712  1.2687811   2.291 0.022521 *
stateMissouri       -4.5010654  1.9061649  -2.361 0.018741 *
stateMontana        -6.0017143  1.9066483  -3.148 0.001782 **
stateNational       -0.0815066  0.6023675  -0.135 0.892442
stateNebraska       -7.9144455  1.9076619  -4.149 4.17e-05 ***
stateNebraska CD-2   4.2176424  1.1795703   3.576 0.000397 ***
stateNevada          1.0493067  0.9342422   1.123 0.262114
stateNew Hampshire   3.2982857  1.9066483   1.730 0.084505 .
stateNew Mexico      4.9448935  1.5837851   3.122 0.001940 **
stateNew York        5.7470479  1.2773507   4.499 9.21e-06 ***
stateNorth Carolina  1.2952212  0.6864443   1.887 0.059982 .
stateOhio           -2.0144661  1.2749875  -1.580 0.114985
statePennsylvania    1.4824831  0.6523022   2.273 0.023632 *
```

```
stateRhode Island    7.3687009  1.9140385   3.850 0.000140 ***
stateTexas          -1.7250989  1.1773303  -1.465 0.143719
stateVirginia        3.7848631  1.1878123   3.186 0.001566 **
stateWisconsin       2.1360233  0.6556746   3.258 0.001230 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.597 on 361 degrees of freedom
Multiple R-squared:  0.4963,    Adjusted R-squared:  0.4572
F-statistic:  12.7 on 28 and 361 DF,  p-value: < 2.2e-16
```

```r
# Plot diagnostic plots for the stepwise model
par(mfrow = c(2, 2))
plot(step_model)
```

```
Warning: not plotting observations with leverage one:
  149, 180, 182, 195, 214
```




```r
par(mfrow = c(1, 1))  # Reset layout
```

Final Models and Predictions for Kamala Harris and Donald Trump

```r
# Final models for Harris and Trump
mlr_harris_model_final <- lm(pct ~ pollscore + log(sample_size) + state, data = harris_data)
mlr_trump_model_final <- lm(pct ~ pollscore + log(sample_size) + state, data = trump_data)

# Predict poll percentages for Kamala Harris
harris_data$predicted_pct_harris <- predict(mlr_harris_model_final, newdata = harris_data)

# Predict poll percentages for Donald Trump
trump_data$predicted_pct_trump <- predict(mlr_trump_model_final, newdata = trump_data)
```

Electoral College Votes Prediction In this step, we aggregate the predictions for each state and determine the winner based on the electoral votes.

```r
# Define electoral votes for each state (and Washington, D.C.)
electoral_votes <- data.frame(
  state = c("Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colorado", "Connecticu
            "Georgia", "Hawaii", "Idaho", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky"
            "Maine CD-1", "Maine CD-2", "Maryland", "Massachusetts", "Michigan", "Minnesota"
            "Missouri", "Montana", "Nebraska", "Nebraska CD-1", "Nebraska CD-2", "Nebraska CD
            "New Hampshire", "New Jersey", "New Mexico", "New York", "North Carolina", "North
            "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island", "South Carolina", "South Da
            "Texas", "Utah", "Vermont", "Virginia", "Washington", "West Virginia", "Wisconsin
            "District of Columbia"),
  electoral_votes = c(9, 3, 11, 6, 55, 9, 7, 3, 29, 16, 4, 4, 20, 11, 6, 6, 8, 8, 2, 1, 1, 10
                      3, 5, 1, 1, 2, 6, 4, 14, 5, 29, 16, 3, 18, 7, 6, 20, 4, 9, 3, 11, 38, 6
)

# Aggregate predictions by state for Harris and Trump
harris_state_avg <- aggregate(predicted_pct_harris ~ state, data = harris_data, FUN = mean)
trump_state_avg <- aggregate(predicted_pct_trump ~ state, data = trump_data, FUN = mean)

# Merge predictions for both candidates
prediction_comparison <- merge(harris_state_avg, trump_state_avg, by = "state", all.x = TRUE)

# Merge the electoral votes with the predictions
prediction_comparison <- merge(prediction_comparison, electoral_votes, by = "state", all.x =

# Determine the winner for each state
prediction_comparison$winner <- ifelse(prediction_comparison$predicted_pct_harris > predictio

# Calculate total electoral votes for Kamala Harris
harris_electoral_votes <- sum(prediction_comparison$electoral_votes[prediction_comparison$win
```

```r
# Calculate total electoral votes for Donald Trump
trump_electoral_votes <- sum(prediction_comparison$electoral_votes[prediction_comparison$winn

# Print the results
print(paste("Harris Electoral Votes:", harris_electoral_votes))
```

```
[1] "Harris Electoral Votes: 216"
```

```r
print(paste("Trump Electoral Votes:", trump_electoral_votes))
```

```
[1] "Trump Electoral Votes: 147"
```

```r
# Determine the predicted winner
if (harris_electoral_votes >= 270) {
  print("Kamala Harris is predicted to win the 2024 election.")
} else if (trump_electoral_votes >= 270) {
  print("Donald Trump is predicted to win the 2024 election.")
} else {
  print("No candidate reached 270 electoral votes.")
}
```

```
[1] "No candidate reached 270 electoral votes."
```

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

```
# #| echo: false
# #| eval: true
# #| warning: false
# #| message: false

# library(rstanarm)

# first_model <-
#   readRDS(file = here::here("models/first_model.rds"))
```

```
# #| echo: false
# #| eval: true
# #| label: tbl-modelresults
# #| tbl-cap: "Explanatory models of flight time based on wing width and wing length"
# #| warning: false

# modelsummary::modelsummary(
#   list(
#     "First model" = first_model
#   ),
#   statistic = "mad",
#   fmt = 2
# )
```

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algo-
rithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.