Politics Are Afoot!

Da Qi Ren

The Setup

There is *a lot* of money that is spent in politics in Presidential election years. So far, estimates have the number at about $11,000,000,000 (11 billion USD). For context, in 2019 Twitter's annual revenue was about $3,500,000,000 (3.5 billion USD).

The work

Install the package, `fec16`.

```
## install.packages('fec16')
```

This package is a compendium of spending and results from the 2016 election cycle. In this dataset are 9 different datasets that cover:

- `candidates`: candidate attributes, like their name, a unique id of the candidate, the election year under consideration, the office they're running for, etc.
- `results_house`: race attributes, like the name of the candidates running in the election, a unique id of the candidate, the number of `general_votes` garnered by each candidate, and other information.
- `campaigns`: financial information for each house & senate campaign. This includes a unique candidate id, the total receipts (how much came in the doors), and total disbursements (the total spent by the campaign), the total contributed by party central committees, and other information.

Your task

Describe the relationship between spending on a candidate's behalf and the votes they receive.

Your work

- We want to keep this work *relatively* constrained, which is why we're providing you with data through the `fec16` package. It is possible to gather all the information from current FEC reports, but it would require you to make a series of API calls that would pull us away from the core modeling tasks that we want you to focus on instead.
- Throughout this assignment, limit yourself to functions that are within the `tidyverse` family of packages: `dplyr`, `ggplot`, `patchwork`, and `magrittr` for wrangling and exploration and `base`, `stats`, `sandwich` and `lmtest` for modeling and testing. You do not *have* to use these packages; but try to limit yourself to using only these.

1. What does the distribution of votes and of spending look like?

1. (3 points) In separate histograms, show both the distribution of votes (measured in `results_house$general_percent` for now) and spending (measured in `ttl_disb`). Use a log transform if appropriate for each visualization. How would you describe what you see in these two plots?

ANSWER:

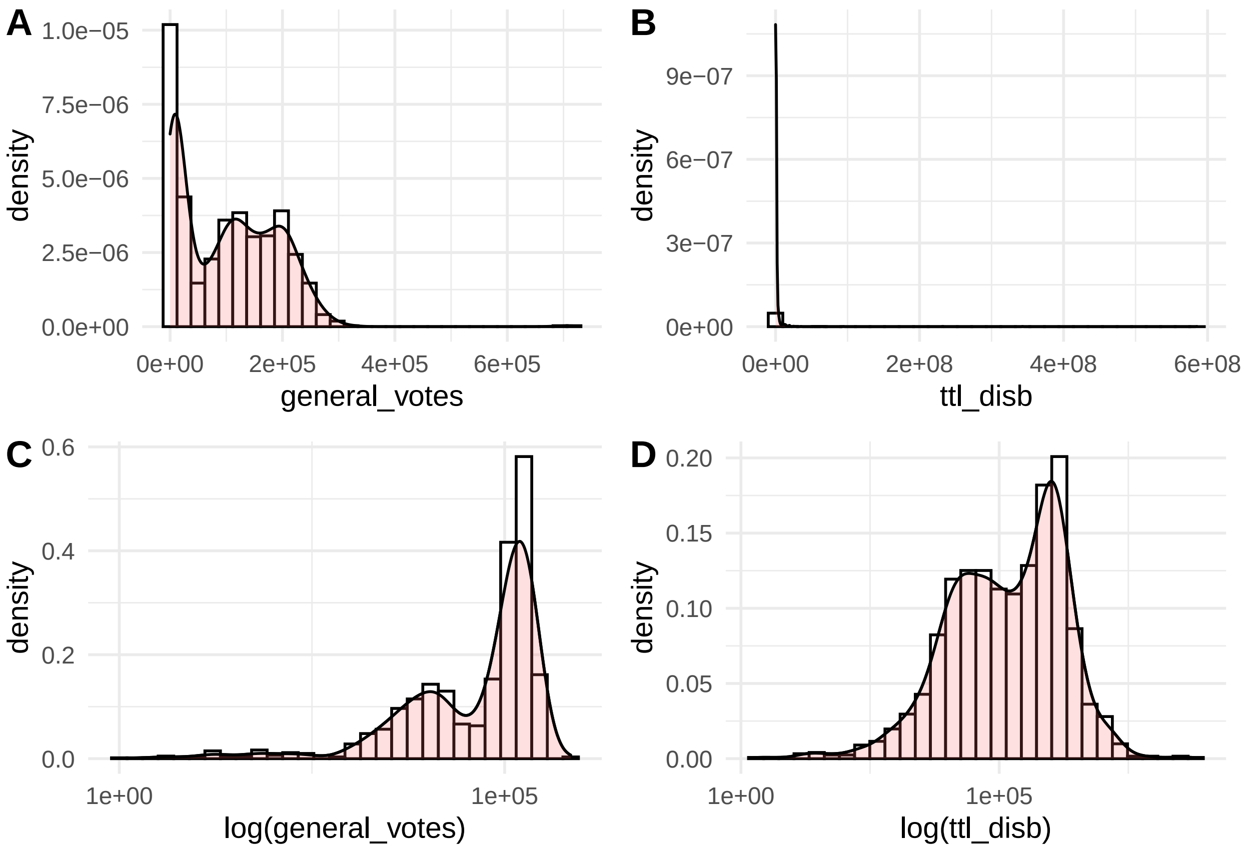
From my observation, the data `general_percent` and `ttl_disb` have the following problems:

- The original data of the 2 variables are not on the same scale (Fig. A-B) .
- Has skewness problems because the curve appears distorted and skewed to the left in a statistical distribution.
- The data are not centered.

At this stage, based on my finding, we need to perform data transforming including scaling, centering and skewness corrections.

I will perform Log transformation first, because log transform makes the data as "normal" as possible so that the statistical analysis results from this data become more valid, the log transformation reduces or removes the skewness of our original data. In detail I choose natural logarithm here for the purposes of linear modeling , i.e., using Log transformation replaces each variable x with a log(x). The results are shown in Fig. C-D, repectivelay. In C and D, after the transfermation, the vurves approcimately follows normal distribution, the graph appears symmetry, there are about as many data values on the left side of the median as on the right side.

I will do other data transformations later in the following questions. Data transformation can make our model working efficiently: distance based models perform well when data is pre-processed and transformed; having all features scaled it speeds up the model; better accuracy and more generalized model.


A
B
C
D
density
general_votes
ttl_disb
log(general_votes)
log(ttl_disb)
1.0e−05
7.5e−06
5.0e−06
2.5e−06
0.0e+00
0e+00
2e+05
4e+05
6e+05
9e−07
6e−07
3e−07
0e+00
2e+08
4e+08
6e+08
0.6
0.4
0.2
0.0
1e+00
1e+05
0.20
0.15
0.10
0.05
0.00

2. Exploring the relationship between spending and votes.

2. (3 points) Create a new dataframe by joining `results_house` and `campaigns` using the `inner_join` function from `dplyr`. (We use the format `package::function` – so `dplyr::inner_join`.)

ANSWER:

Done the creation of new dataframe by joining `results_house` and `campaigns` using the `inner_join` function from `dplyr`. The new data frame is named "d1". A discription of "d1" is as the follows:

```
d1 <- dplyr::inner_join(results_house, campaigns, by = NULL)
```

```
## Joining, by = "cand_id"
```

```
nrow(d1)
```

```
## [1] 1342
```

```
summary(d1)
```

```
##     state            district_id          cand_id           incumbent
##  Length:1342        Length:1342        Length:1342        Mode :logical
##  Class :character   Class :character   Class :character   FALSE:895
##  Mode  :character   Mode  :character   Mode  :character   TRUE :447
##
##
##
##
##     party            primary_votes    primary_percent    runoff_votes
##  Length:1342        Min.   :     1   Min.   :0.00015   Min.   : 1096
##  Class :character   1st Qu.:  8650   1st Qu.:0.19158   1st Qu.: 1464
##  Mode  :character   Median : 21299   Median :0.42257   Median : 8206
##                     Mean   : 32227   Mean   :0.48844   Mean   :11274
##                     3rd Qu.: 45638   3rd Qu.:0.78382   3rd Qu.:20082
##                     Max.   :326988   Max.   :1.00000   Max.   :25322
##                     NA's   :291      NA's   :292       NA's   :1330
##  runoff_percent   general_votes    general_percent     won
##  Min.   :0.3427   Min.   :    55   Min.   :0.0000   Mode :logical
##  1st Qu.:0.4624   1st Qu.: 88229   1st Qu.:0.3087   FALSE:850
##  Median :0.5000   Median :142597   Median :0.4773   TRUE :492
##  Mean   :0.5000   Mean   :136932   Mean   :0.4597
##  3rd Qu.:0.5376   3rd Qu.:198290   3rd Qu.:0.6406
##  Max.   :0.6573   Max.   :718591   Max.   :1.0000
##  NA's   :1330     NA's   :462      NA's   :463
##   footnotes          cand_name           cand_ici             pty_cd
##  Length:1342        Length:1342        Length:1342        Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :1.607
##                                                           3rd Qu.:2.000
##                                                           Max.   :3.000
##
```

```
##  cand_pty_affiliation  ttl_receipts      trans_from_auth      ttl_disb
##  Length:1342          Min.   :       0   Min.   :       0   Min.   :       0
##  Class :character     1st Qu.:   46612   1st Qu.:       0   1st Qu.:   46147
##  Mode  :character     Median :  398962   Median :       0   Median :  379570
##                       Mean   :  883177   Mean   :   26408   Mean   :  814754
##                       3rd Qu.: 1290266   3rd Qu.:       0   3rd Qu.: 1154148
##                       Max.   :19852221   Max.   :12374657   Max.   :13433669
##
##  trans_to_auth      coh_bop           coh_cop         cand_contrib
##  Min.   :     0   Min.   : -18681   Min.   : -32074   Min.   :       0
##  1st Qu.:     0   1st Qu.:      0   1st Qu.:      0   1st Qu.:       0
##  Median :     0   Median :      0   Median :   3881   Median :       0
##  Mean   :  7577   Mean   : 150271   Mean   : 218929   Mean   :   21879
##  3rd Qu.:     0   3rd Qu.:  85884   3rd Qu.: 170548   3rd Qu.:    1000
##  Max.   :766500   Max.   :3750024   Max.   :9098873   Max.   :13414225
##
##    cand_loans       other_loans     cand_loan_repay   other_loan_repay
##  Min.   :      0   Min.   :     0   Min.   :      0   Min.   :     0.0
##  1st Qu.:      0   1st Qu.:     0   1st Qu.:      0   1st Qu.:     0.0
##  Median :      0   Median :     0   Median :      0   Median :     0.0
##  Mean   :  56809   Mean   :  1049   Mean   :  12579   Mean   :   638.7
##  3rd Qu.:   9000   3rd Qu.:     0   3rd Qu.:      0   3rd Qu.:     0.0
##  Max.   :8050000   Max.   :350000   Max.   :1655854   Max.   :350000.0
##
##  debts_owed_by     ttl_indiv_contrib cand_office_st     cand_office_district
##  Min.   :  -1786   Min.   :      0   Length:1342        Length:1342
##  1st Qu.:      0   1st Qu.:  21310   Class :character   Class :character
##  Median :      0   Median : 207337   Mode  :character   Mode  :character
##  Mean   :  42528   Mean   : 464597
##  3rd Qu.:  12903   3rd Qu.: 638629
##  Max.   :2795000   Max.   :5975190
##
##  other_pol_cmte_contrib pol_pty_contrib   cvg_end_dt          indiv_refunds
##  Min.   :      0        Min.   :    0   Min.   :2015-08-10   Min.   : -1150
##  1st Qu.:      0        1st Qu.:    0   1st Qu.:2016-12-31   1st Qu.:     0
##  Median :  13700        Median :    0   Median :2016-12-31   Median :   200
##  Mean   : 305670        Mean   : 1230   Mean   :2016-11-30   Mean   :  6617
##  3rd Qu.: 506471        3rd Qu.:  150   3rd Qu.:2016-12-31   3rd Qu.:  5400
##  Max.   :3279747        Max.   :25400   Max.   :2017-01-31   Max.   :227497
##
##   cmte_refunds
##  Min.   :     0
##  1st Qu.:     0
##  Median :     0
##  Mean   :  1093
##  3rd Qu.:   250
##  Max.   :104758
##
```

3. (3 points) Produce a scatter plot of `general_votes` on the y-axis and `ttl_disb` on the x-axis. What do you observe about the shape of the joint distribution?

ANSWER:

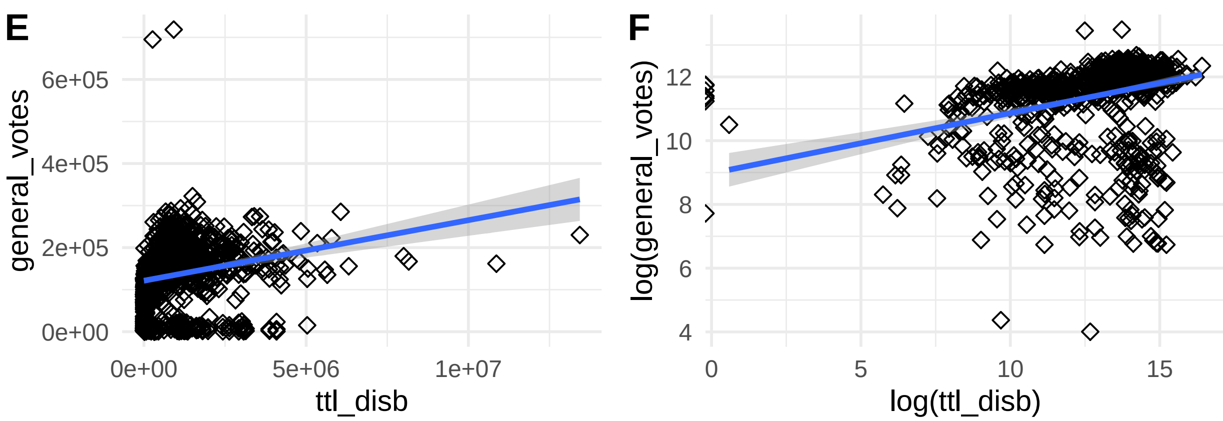
The scatter plot of `general_votes` on the y-axis and `ttl_disb` on the x-axis is shown below Fig.E. I also made a scatter plot using y = log(general_votes) and x= log(ttl_disb), as shown in Fig.F.

In general, a x-y scatter graph displays and compares values to show the numerical distribution of variables in a rectangular coordinate system. A two-dimensional scatter chart can show the data analysis of two variables to provide the relationship and correlation between the two. Scatter plots can provide three types of key information:

- Whether there is a quantitative correlation trend between variables;
- If there is a correlation trend, is it linear or non-linear;
- Observe whether there are outliers and analyze The influence of these outliers on the modeling analysis.

However, I couldn't find obvious correlation between variables since most of them look randomly distributed on the scatter plot. If there is a certain correlation, then most of the data points will be relatively dense and present in a certain trend, however I cannot figure it out it by simple observation.

By observing the distribution of data points on the scatter plot, I found there are some outliers.



4. (3 points) Create a new variable to indicate whether each individual is a “Democrat”, “Republican” or “Other Party”.

- Here’s an example of how you might use `mutate` and `case_when` together to create a variable.

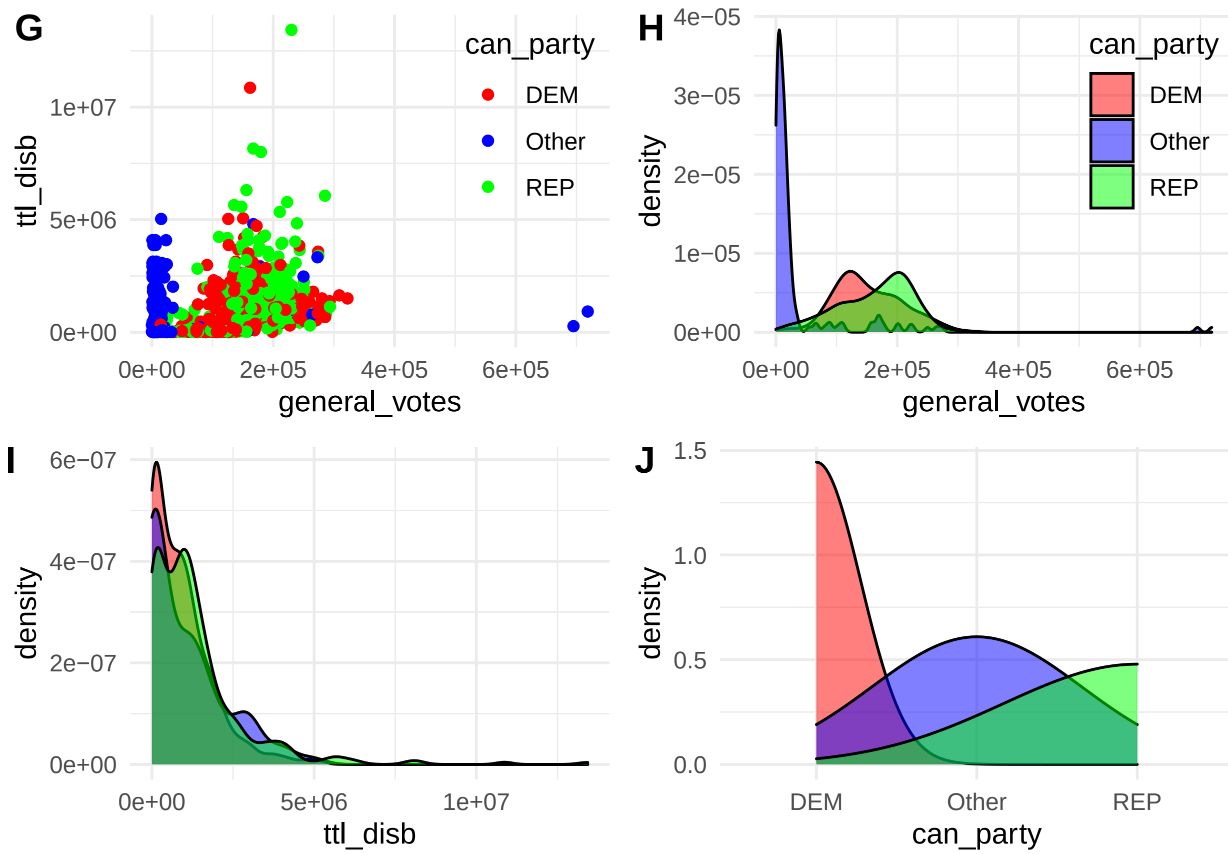
```
starwars %>%
  select(name:mass, gender, species) %>%
  mutate(
  type = case_when(
    height > 200 | mass > 200 ~ "large",
    species == "Droid"        ~ "robot",
    TRUE                      ~ "other"
    )
  )
```

Once you’ve produced the new variable, plot your scatter plot again, but this time adding an argument into the `aes()` function that colors the points by party membership. What do you observe about the distribution of all three variables?

ANSWER:

The new variable has been produced, the new data frame is named “d2”, a discription is as the follows :

```
##   can_party          general_votes       ttl_disb
##  Length:880          Min.   :    55   Min.   :       0
##  Class :character   1st Qu.: 88229   1st Qu.:  102276
##  Mode  :character   Median :142597   Median :  830659
##                      Mean   :136932   Mean   : 1084565
##                      3rd Qu.:198290   3rd Qu.: 1527533
##                      Max.   :718591   Max.   :13433669
```



From my observation in Fig H-J, the variable general_percent,ttl_disb and can_party have the following properties:

- The distribution of each of the three variables (i.e. can_party, ttl_disp, general_vote) are a combination of 3 different curves that are approximately following normal distributions.
- For each variable, the 3 curves in different color clustered by the 3 (i.e. DEM , REP, and Other) parties.
- Among the total 9 curves, each of the curves appears symmetry, there are about as many data values on the left side of the median as on the right side.
- Each of the curves has skewness problems because the curve appears distorted or skewed to the left or right in a statistical distribution.
- The data in each curve are not centered.

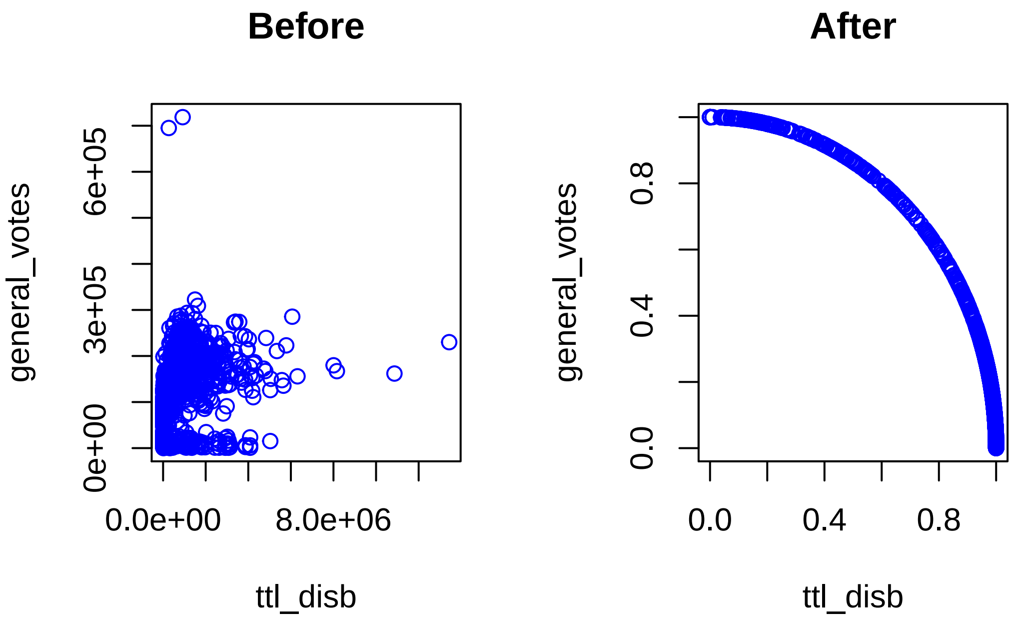
At this stage, based on my finding, the following decisions are made:

- A linear model can be created and fit the relationship between the general_votes and ttl_disb and can_party.
- Detailed analysis and pre-processing need to be done to the data using maths.
- further data transformations have to be performed.

Next, I will do the data pre-processing and model creation.

Produce a Descriptive Model

5. (5 Points) Given your observations, produce a linear model that you think does a good job at describing the relationship between candidate spending and votes they receive. You should decide what transformation to apply to spending (if any), what transformation to apply to votes (if any) and also how to include the party affiliation.




```
##
## Call:
## lm(formula = d2$general_votes ~ d2$ttl_disb + d2$can_party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -162812  -50839    -463   37128  645725
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.634e+05  4.080e+03  40.061  < 2e-16 ***
## d2$ttl_disb   1.163e-02  1.864e-03   6.238 6.88e-10 ***
## d2$can_party -5.062e+04  3.213e+03 -15.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69240 on 877 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.2585
## F-statistic: 154.2 on 2 and 877 DF,  p-value: < 2.2e-16
```

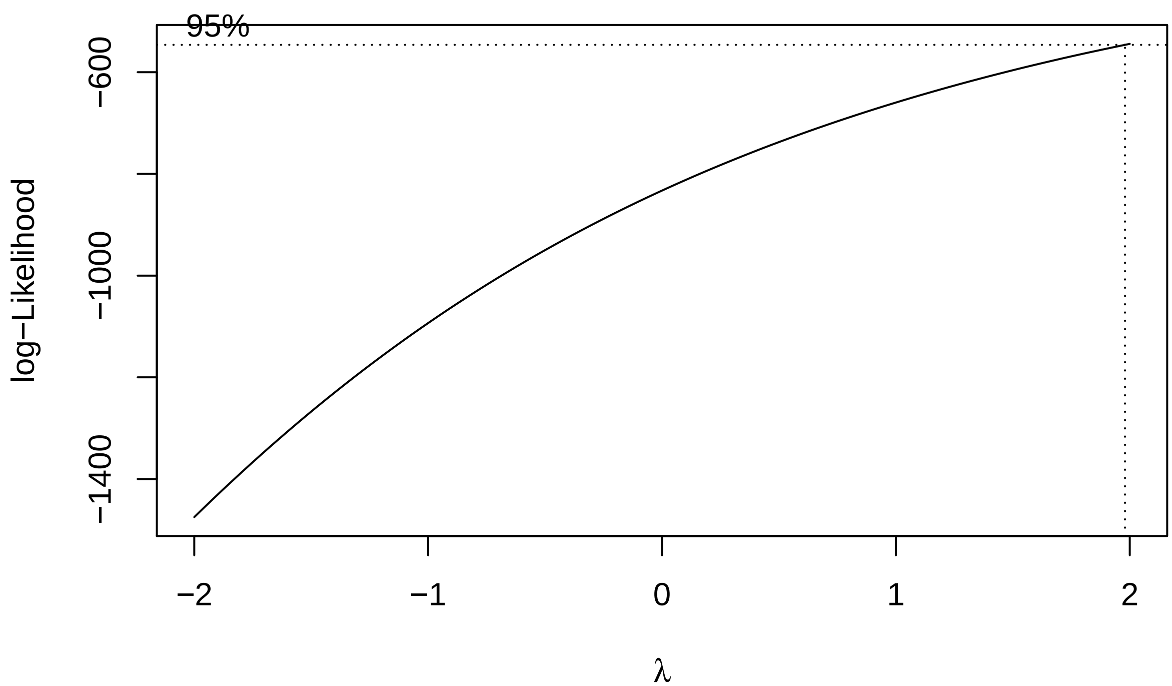
```
##
## Call:
## lm(formula = d2$csvotes ~ d2$csdisb + d2$csparty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.0249 -0.6323 -0.0058  0.4618  8.0309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.857e-16  2.903e-02   0.000        1
## d2$csdisb    1.820e-01  2.917e-02   6.238 6.88e-10 ***
## d2$csparty  -4.597e-01  2.917e-02 -15.756  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8611 on 877 degrees of freedom
## Multiple R-squared:  0.2602, Adjusted R-squared:  0.2585
## F-statistic: 154.2 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = d2$novotes ~ d2$nodisb + d2$noparty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25036 -0.09079 -0.01375  0.08107  0.24505
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.25036    0.01271   98.38   <2e-16 ***
## d2$nodisb    -1.09452    0.01426  -76.76   <2e-16 ***
## d2$noparty  -80.80898   79.99695   -1.01    0.313
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.117 on 877 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8774
## F-statistic:  3146 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = d2$logvotes ~ d2$logdisb + d2$logparty)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3555 -0.1927  0.0741  0.2940  4.1067
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.60108    0.16135  65.704  < 2e-16 ***
## d2$logdisb   0.09767    0.01226   7.968 5.01e-15 ***
## d2$logparty -3.57205    0.10352 -34.507  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.809 on 877 degrees of freedom
## Multiple R-squared:  0.6044, Adjusted R-squared:  0.6035
## F-statistic: 669.9 on 2 and 877 DF,  p-value: < 2.2e-16
```

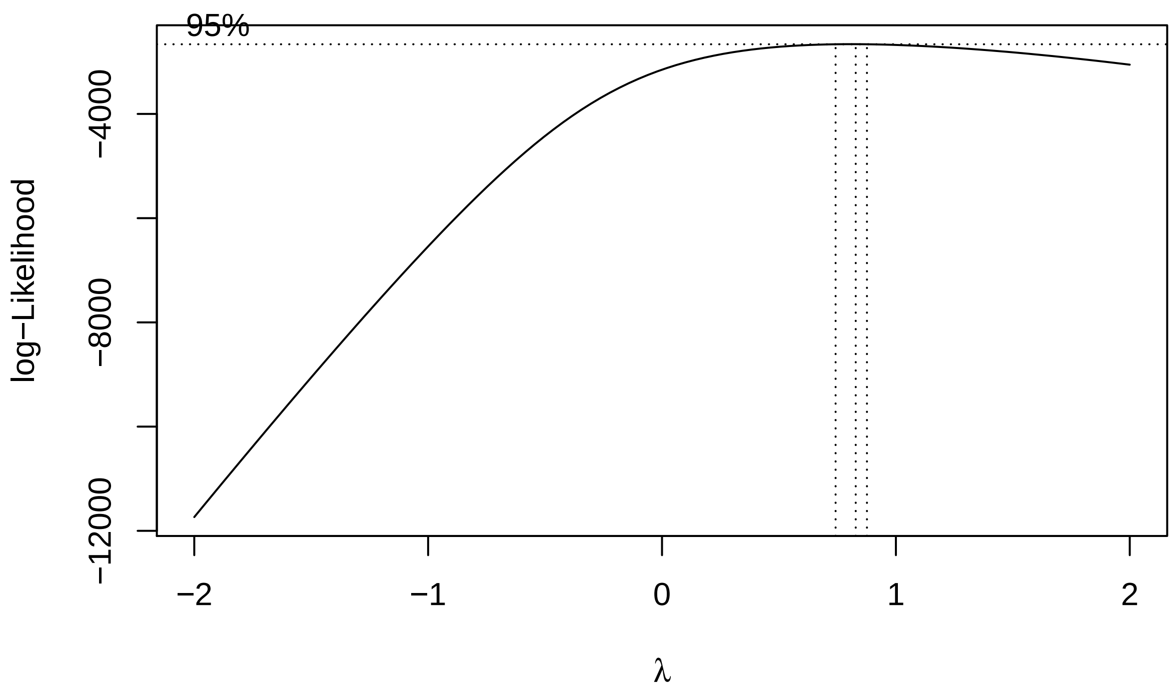
```
##
## Call:
## lm(formula = d2$general_votes ~ d2$logdisb + d2$can_party)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -164084  -42521    2037   33117  627966
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20183.4    13565.1   1.488    0.137
## d2$logdisb    11935.7      999.7  11.939   <2e-16 ***
## d2$can_party -46716.3     3070.3 -15.215   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65620 on 877 degrees of freedom
## Multiple R-squared:  0.3354, Adjusted R-squared:  0.3339
## F-statistic: 221.3 on 2 and 877 DF,  p-value: < 2.2e-16
```




```
##         lambda       lik
## [1,] -2.000000 -1474.865
## [2,] -1.959596 -1456.833
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.727   -6.848    1.412    8.771  119.849
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.737      4.067  29.192   <2e-16 ***
## logdisb        3.029      0.309   9.802   <2e-16 ***
## logparty     -91.757      2.610 -35.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 877 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:   0.62
## F-statistic: 718.1 on 2 and 877 DF,  p-value: < 2.2e-16
```




```
##         lambda       lik
## [1,] -2.000000 -11735.61
## [2,] -1.959596 -11515.25
```

```
##
## Call:
## lm(formula = lam2votes ~ nodisb + noparty, data = d2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25036 -0.09079 -0.01375  0.08107  0.24505
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.25036    0.01271   19.70   <2e-16 ***
## nodisb       -1.09452    0.01426  -76.76   <2e-16 ***
## noparty     -80.80898   79.99695   -1.01    0.313
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.117 on 877 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8774
## F-statistic:  3146 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.727   -6.848    1.412    8.771  119.849
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  118.737      4.067  29.192   <2e-16 ***
## logdisb        3.029      0.309   9.802   <2e-16 ***
## logparty     -91.757      2.610 -35.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.39 on 877 degrees of freedom
## Multiple R-squared:  0.6209, Adjusted R-squared:   0.62
## F-statistic: 718.1 on 2 and 877 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = lamvotes ~ logdisb + logparty, weights = 1/abs(e))
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -10.184  -2.701   1.057   2.873  10.958
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118.10513    0.82784  142.67   <2e-16 ***
## logdisb       3.09811    0.06612   46.85   <2e-16 ***
## logparty    -92.24965    0.34882 -264.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.571 on 877 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9881
## F-statistic: 3.635e+04 on 2 and 877 DF,  p-value: < 2.2e-16
```

6. (3 points) Interpret the model coefficients you estimate.

- Tasks to keep in mind as you're writing about your model:
  - At the time that you're writing and interpreting your regression coefficients you'll be *deep* in the analysis. Nobody will know more about the data than you do, at that point. *So, although it will feel tedious, be descriptive and thorough in describing your observations.*
  - It can be hard to strike the balance between: on the one hand, writing enough of the technical underpinnings to know that your model meets the assumptions that it must; and, on the other hand, writing little enough about the model assumptions that the implications of the model can still be clear. We're starting this practice now, so that by the end of Lab 2 you will have had several chances to strike this balance.

ANSWER

Interpreting a Coefficient as a Rate of Change in Y Instead of as a Rate of Change in the Conditional Mean of Y.

As pointed out in the discussion of overfitting, the computed regression equation estimates the true conditional mean function. How well it estimates the behavior of actual values of the random variable depends on the variability of the response variable Y. Thus, interpreting the computed coefficients in terms of the response variable is often misleading.

Illustration: In the graph shown below, the data are marked in green, the true line of conditional means is in violet, and the fitted (computed) regression line is in blue. Note that the fitted regression line is close to the true line of conditional means. The equation of the fitted regression line is (with coefficients rounded to a reasonable degree) ŷ = 0.56 + 2.18x.1 Thus it is accurate to say, "For each change of one unit in x, the average change in the mean of Y is about 2.18 units." It is not accurate to say, "For each change of one unit in x, Y changes about 2.18 units." For example, we can see from the graph that when x is 2, Y might be anywhere between a little below 4 to a little above 5.5; when x is 3, Y might be anywhere from a little more than 5.5 to a little more than 9. So when going from x = 2 to x = 3, the change in Y might be almost zero, or it might be as large as 5.5 units.

Interpreting a coefficient as a rate of change in Y instead of as a rate of change in the conditional mean of Y.

2. Not taking confidence intervals for coefficients into account.

Even when a regression coefficient is (correctly) interpreted as a rate of change of a conditional mean (rather than a rate of change of the response variable), it is important to take into account the uncertainty in the estimation of the regression coefficient. To illustrate, in the example used in item 1 above, the computed regression line has equation ŷ = 0.56 + 2.18x. However, a 95% confidence interval for the slope is (1.80, 2.56). So saying, "The rate of change of the conditional mean of Y with respect to x is estimated to be between 1.80 and 2.56" is usually1 preferable to saying, "The rate of change of the conditional mean Y with respect to x is about 2.18."

3. Interpreting a coefficient that is not statistically significant.2

Interpretations of results that are not statistically significant are made surprisingly often. If the t-test for a regression coefficient is not statistically significant, it is not appropriate to interpret the coefficient. A better alternative might be to say, "No statistically significant linear dependence of the mean of Y on x was detected.

4. Interpreting coefficients in multiple regression with the same language used for a slope in simple linear regression.

Even when there is an exact linear dependence of one variable on two others, the interpretation of coefficients is not as simple as for a slope with one dependent variable.

Example: If y = 1 + 2x1 + 3x2, it is not accurate to say "For each change of 1 unit in x1, y changes 2 units". What is correct is to say, "If x2 is fixed, then for each change of 1 unit in x1, y changes 2 units."

Similarly, if the computed regression line is $\hat{y}$ = 1 + 2x1 + 3x2, with confidence interval (1.5, 2.5), then a correct interpretation would be, "The estimated rate of change of the conditional mean of Y with respect to x1, when x2 is fixed, is between 1.5 and 2.5 units."

For more on interpreting coefficients in multiple regression, see Section 4.3 (pp 161-175) of Ryan3.

5. Multiple inference on coefficients.

Chapter 1: Sources & Types of Regression Output

Regression analysis can be performed on a variety of software today. The ubiquitous Microsoft Excel is still by far the most popular tool. A variety of other free and paid tools are available to run regression analysis. Some of these include SPSS, SAS, R, Python and JMP, etc.

Each of these tools presents the regression analysis output data in different ways. However, all of these tools provide essentially the same data. We present below the regression output from some of the tools mentioned above.

The raw data is available on the book's webpage here. Please feel free to play with it live and see the impact it has on the regression equation and the corresponding chart. 1.1 Microsoft Excel Output Regression Analysis Output in Microsoft Excel 1.2 R Programming Output Regression Analysis Output in R

Note that in all these cases, the regression analysis output provides essentially the same information although it is presented in different formats or designs.

Every number in the regression output indicates something. We will address only the most frequently used numbers in this book.

================ Chapter 2: The Big Picture Understanding the Model

The first set of numbers my eyes wander to are at the top of the regression output in Microsoft Excel under the heading Regression Statistics.

Regression Statistics

This data is presented in the last few rows of the regression output in R. This set of data gives you the big picture about your regression output. It allows you to answer questions such as: How good is your model? What percentage of the variation is explained by the variables included?

============

2.1 The Multiple R

The multiple R is the absolute value of the correlation coefficient of the two variables (X and Y) being evaluated. The correlation coefficient indicates how closely two variables move in tandem with each other. It assumes that the relationship is linear and so measures the linear relationship between the two variables X and Y.

The correlation coefficient has a value between and . A correlation coefficient of indicates that the variables move in perfect tandem and in the same direction. A correlation coefficient of 0 indicates that there is no relationship between the variables. A correlation coefficient of indicates that the variables move in perfect tandem but in the OPPOSITE direction.

However, since the multiple R is the absolute value of the correlation coefficient, we do not get to know if the correlation is positive or negative! This means that we do not see the direction of the relationship and only know the strength of the relationship.

The correlation coefficient is also referred to as the Pearson correlation coefficient or the Pearson's r.

The Multiple R in our example indicates that there is a strong correlation between the amount spent on TV ads and sales. As indicated above, the Multiple R will not tell us if the correlation is positive or negative.

========

2.2 R-Squared or Multiple R-Squared

The R-Squared (in Microsoft Excel) or Multiple R-Squared (in R) indicates how well the model or regression line "fits" the data. It indicates the proportion of variance in the dependent variable (Y) that is explained by the independent variable (X).

We know a variable could be impacted by one or more factors. The R-Squared indicates the percentage of variation in the dependent variable that is explained by the independent variables.

In our example, we know that the unit sales of a product will be influenced by a variety of factors such as price, competitors' actions, economy, etc. and not just by the advertisement expenditure. When we run a regression with sales as the dependent Y variable and only advertisement expenditure as the independent X variable, the R-square indicates the percentage of variation in unit sales that is explained by the advertisement expenditure. It tells you the percentage of change in sales that is caused by varying the advertisement expenditure. This also means that we can compute the percentage of variation that is explained by factors other than advertisement expenditure such as the economy, competition, price, etc. The percentage of variation that is explained by factors other than advertisement expenditure will be 100%-R-square.

Our regression output indicates that 81.48% of variation in unit sales is explained by the advertisement budget. And 18.52% (100%-81.48%) of the variation is caused by factors other than advertisement expenditure.

(Also note that as the name suggests, the R-square is equal to the square of the multiple R!) 2.3 Adjusted R-Squared

===========

Adjusted R-Squared is used only when analyzing multiple regression output and ignored when analyzing simple linear regression output. When we have more than one independent variable in our analysis, the computation process inflates the R-squared. As the name indicates, the Adjusted R-Squared is the R-Square adjusted for this inflation when performing multiple regression.

The interpretation of the Adjusted R-Squared is similar to the R-square and used only when analyzing multiple regression output. 2.4 The Standard Error

The standard error in the regression output is a very important number to understand when interpreting regression data. The standard error is a measure of the precision of the model. It reflects the average error of the regression model. In other words, if we were using the regression model to predict or estimate the dependent variable or variable of interest, the standard error shows you how wrong you could be if you used the regression model to make predictions. As the standard error reflects how wrong you could be, we want the standard error to be as small as possible.

The standard error is used to help you get a confidence interval for your predicted values. 2.5 Significance F

The simplest way to understand the significance F is to think of it as the probability that our regression model is wrong and needs to be discarded!! The significance F gives you the probability that the model is wrong. We want the significance F or the probability of being wrong to be as small as possible.

Significance F: Smaller is better…. Significance F in Regression Output

We can see that the Significance F is very small in our example. We usually establish a significance level and use it as the cutoff point in evaluating the model. Commonly used significance levels are 1%, 5% or 10%.

Statistically speaking, the significance F is the probability that the null hypothesis in our regression model cannot be rejected. In other words, it indicates the probability that all the coefficients in our regression output are actually zero! The significance F is computed from the F value (found to the left of the significance F in Microsoft Excel's output). The F value is a value similar to the z value, t value, etc. It is a ratio computed by dividing the mean regression sum of squares by the mean error sum of squares. The F value ranges from zero to a very large number.

Note that the significance F is similar in interpretation to the P value discussed later a later section. The key difference is that the significance F applies to the entire model as a whole whereas the P value will be applied only to each corresponding coefficient.