Assignment 9 (Extra Credit)

36-350, Data Mining

Due Monday, 15 December, at 10:30 am

1. *Can you predict all of the people some of the time?*

   (a) *Calculate the mean squared error of each forecaster's predictions for the vote share. What is the ranking of forecasters?*

   ANSWER: The wording unfortunately is ambiguous; *whose* vote share? Or the difference in vote shares? I meant the latter (the "margin" in the electoral sense). Start with the error for 538.com:

```
> mean(votes[,"Obama.margin"] - votes[,"Projected.Margin.538"])^2
[1] 37.27839
```

   Note the units here: the observations are recorded in percent, so this has units of percent *squared*. If I want to go back to percent I need to take the square root (root-mean-square or RMS error):

```
> sqrt(mean(votes[,"Obama.margin"] - votes[,"Projected.Margin.538"])^2)
[1] 6.105603
```

   Similarly for the others:

```
> sqrt(mean(votes[,"Obama.margin"] - votes[,"Projected.Margin.EV"])^2)
[1] 5.379751
> sqrt(mean((votes[,"Obama.margin"] - votes[,"Projected.Margin.Pollster"])^2))
[1] 5.205717
```

   So Pollster.com had the lowest mean-squared error, followed by E-V, followed by 538.com.

   Alternately, we can *check* the computation of the columns which supposedly give us the magnitude (absolute value) of the errors:

```
> summary(abs(votes[,"Obama.margin"] - votes[,"Projected.Margin.538"])
          - votes[,"Magnitude.actual.margin.minus.538"])
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-7.105e-15 -4.441e-16  0.000e+00 -2.705e-16  2.082e-16  2.276e-15
```

   These are all well below the level we'd expect from numerical round-off, especially if someone was using something like Excel[1] Since the this checks out, we could do

[1] Don't use Excel to do statistics. Please.

```
> sqrt(mean((votes[,"Magnitude.actual.margin.minus.Pollster"])^2))
[1] 5.205717
```

and get the same answer.

(b) *Calculate the median absolute error. Does the ranking change?*

ANSWER:

```
> median(votes[,"Magnitude.actual.margin.minus.538"])
[1] 2.4
> median(votes[,"Magnitude.actual.margin.minus.EV"])
[1] 2.96
> median(votes[,"Magnitude.actual.margin.minus.Pollster"])
[1] 2.32
```

So the ranking has changed: Pollster still has the lowest error, but 538 is now better than E-V.

(c) *Calculate the mis-classification rate, i.e., the fraction of states where the candidate predicted to get a majority actually got only a minority. What is the ranking?*

ANSWER: As usual, the classification depends only on the sign of the margin.

```
> sum(sign(votes[,"Obama.margin"]) != sign(votes[,"Projected.Margin.538"]))/51
[1] 0.01960784
> sum(sign(votes[,"Obama.margin"]) != sign(votes[,"Projected.Margin.EV"]))/51
[1] 0.03921569
> sum(sign(votes[,"Obama.margin"]) != sign(votes[,"Projected.Margin.Pollster"]))/51
[1] 0.03921569
```

In other words, 538 got one state wrong (Indiana), while EV and Pollster both got two wrong (Indiana and Missouri). Thus 538 does best by this, followed by the other two.

(d) *Calculate the weighted mean squared error, with each state weighted by population. What happens to the ranking?*

ANSWER: Nobody looked up population figures; instead everyone used total votes cast, which is fair enough. The first three columns of the matrix give the number of votes cast for Obama, McCain and for minor candidates, respectively. We'll add those up and use them to make a weighted average. These commands give the MSE and the RMS error; the latter is easier to interpret.

```
> total.votes = rowSums(votes[,1:3])
> vote.weights = total.votes/sum(total.votes)
> sum((votes[,"Magnitude.actual.margin.minus.538"])^2 * vote.weights)
[1] 10.83356
> sum((votes[,"Magnitude.actual.margin.minus.EV"])^2 * vote.weights)
[1] 14.28095
```

```
> sum((votes[,"Magnitude.actual.margin.minus.Pollster"])^2 * vote.weights)
[1] 10.89593
```

By running `sqrt`, we get that the RMS errors are (in the same order) $3.29, 3.78, 3.30$. So the ranking is 538, followed by EV, followed by Pollster — exactly the opposite of the ranking by unweighted MSE.

(Weighting the squared error by the total number of votes cast would be appropriate if the variance was proportional to the number of votes, or $\sigma \propto \sqrt{n}$, as in a binomial distribution. Since voters are *not* independently-selected random samples from an underlying population, it seems unlikely that the right model really is a binomial. But a better model isn't obviously available. Also, notice that the weighted RMS errors are much better than the unweighted ones.)

(e) *Plot error versus predicted margin for each forecaster, and describe any patterns.*

ANSWER: Here's the code I used to plot the error versus prediction for 538.

```
plot(votes[,"Projected.Margin.538"],
     votes[,"Obama.margin"]-votes[,"Projected.Margin.538"],
     xlab="Prediction: 538",ylab="Error: Actual minus Prediction")
abline(h=0,col="grey"); abline(v=0,col="grey")
lines(ksmooth(votes[,"Projected.Margin.538"],
      votes[,"Obama.margin"]-votes[,"Projected.Margin.538"],
      "normal",bandwidth=8),col="grey")
```

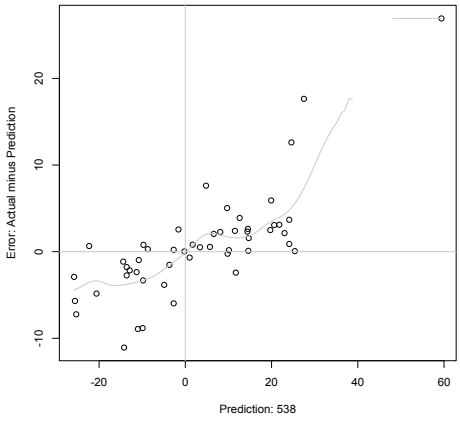
The extra commands add horizontal and vertical axes (through the origin), and a smoothing curve (just to guide the eye). Most of the points fall either in the upper-right or the bottom-left quadrants, and the trend curve is roughly diagonal, all implying that 538's errors are bigger when it makes more extreme predictions, and that the error is in the same direction as the prediction — i.e., the systematic problem with 538's predictions are that they aren't extreme enough, that when it predicts one candidate will lead by 20 points it should really predict they'd lead by (roughly) 25.

Repeating this for EV and Pollster (code omitted), we see that Pollster has a similar problem (though in a less extreme form), while EV is much more balanced in its errors.

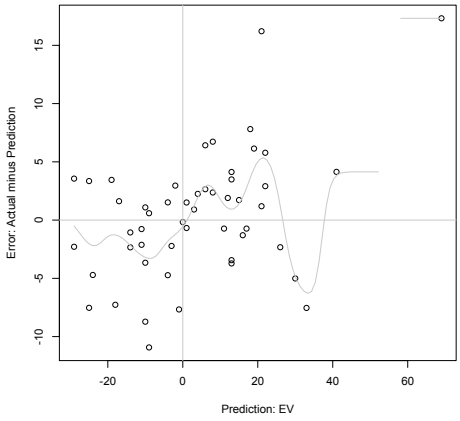
(f) *Plot error versus* actual *margins, and describe any patterns.*

ANSWER: Almost the same code will do.

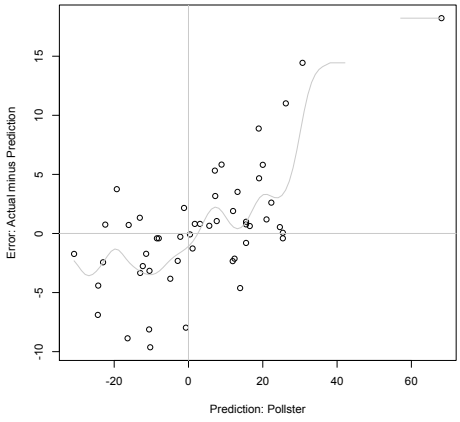
```
plot(votes[,"Obama.margin"],
     votes[,"Obama.margin"]-votes[,"Projected.Margin.538"],
     xlab="Actual Margin",ylab="Error: Actual minus Prediction",
     main = "538: error vs. actual margin")
abline(h=0,col="grey"); abline(v=0,col="grey")
lines(ksmooth(votes[,"Obama.margin"],
```



EV: error vs. prediction



Pollster: error vs. prediction



```
votes[,"Obama.margin"]-votes[,"Projected.Margin.538"],
"normal",bandwidth=8),col="grey")
```

Unsurprisingly, the systematic trends are similar, though perhaps less pronounced.

(g) *Plot error versus the total number of votes cast, by state. Describe any patterns.*

ANSWER: We want code like this.

```
plot(total.votes,votes[,"Magnitude.actual.margin.minus.538"],
     xlab="Total votes",ylab="Error magnitude",
     main="538: size of error vs total votes cast",log="x")
```

This plots the magnitude of the error against the number of votes cast, with the latter on a log scale for clarity. Clearly, the size of the error tends to shrink as the state gets larger. The same is true of the other two, but less dramatically.

(h) *Which forecaster would you trust in 2012? (Explain.)*

ANSWER: All three perform very similarly. Pollster has the lowest MSE and MAE, 538 has the lowest weighted error and misclassification rate, and EV is less systematic in its mistakes, with error scores very close to the other two.

2. *Rates and costs*

(a) *Patient who do not have X but are falsely diagnosed with it must undergo painful and embarrassing follow-up tests, equivalent to a cost to them of $5,000. Patients who have X but are not diagnosed with it die, which for the purposes of this problem is worth $10,000,000 to avoid. Find the expected cost to the patient of: taking no test, taking the old test, and taking the new test, all as functions of p.*

ANSWER:

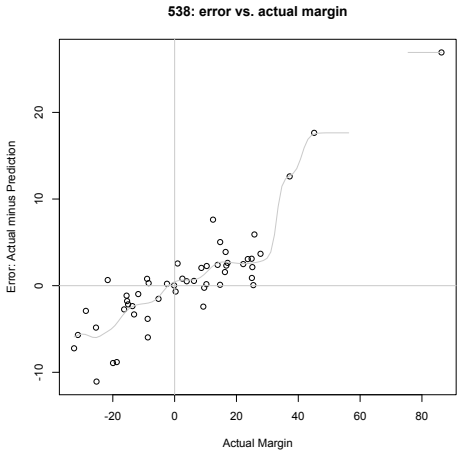
$$\begin{aligned}
\mathbf{E}\left[\text{no test}\right] &= p \times 10^7 \\
\mathbf{E}\left[\text{old test}\right] &= p \times 0.1 \times 10^7 = 10^6 p \\
\mathbf{E}\left[\text{new test}\right] &= (1-p) \times 0.01 \times 5 \times 10^3 + p \times 0.01 \times 10^7 \\
&= 50 + (10^5 - 50)p
\end{aligned}$$

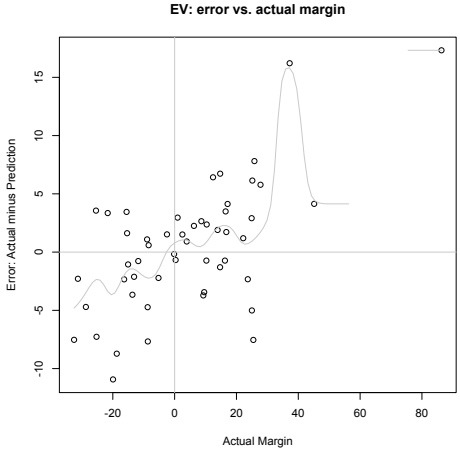
(b) *When, as a function of p, would the patient want to take the new test, the old test, or no test at all?*

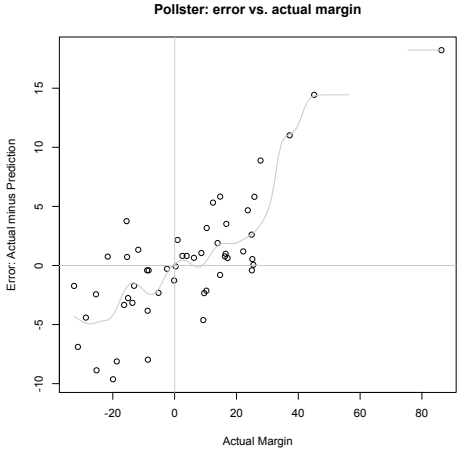
ANSWER: There is no value of p at which a patient would prefer no test to the old test (since $10^7 > 10^6$).

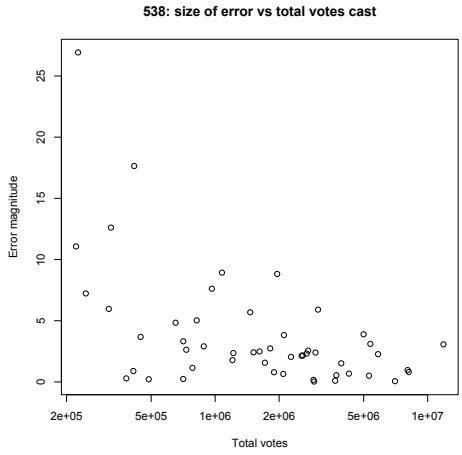
Patients will prefer the old test to the new test when the former's expected cost is lower, i.e., when

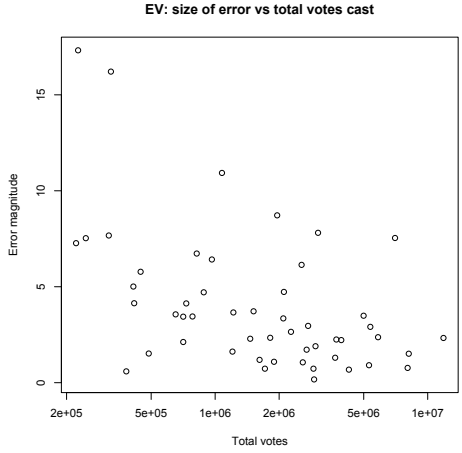
$$\begin{aligned}
10^6 p &< 50 + (10^5 - 50)p \\
(10^6 - 10^5 + 50)p &< 50 \\
p &< 5.6 \times 10^{-5}
\end{aligned}$$

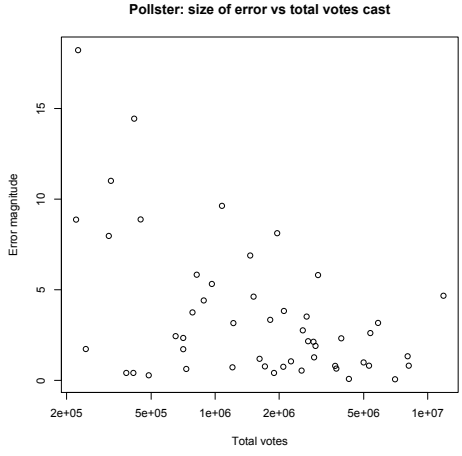

538: error vs. actual margin
Error: Actual minus Prediction
Actual Margin
-20
0
20
40
60
80
-10
0
10
20


EV: error vs. actual margin
Error: Actual minus Prediction
Actual Margin
-20
0
20
40
60
80
-10
-5
0
5
10
15


Pollster: error vs. actual margin
Error: Actual minus Prediction
Actual Margin
-20
0
20
40
60
80
-10
-5
0
5
10
15


538: size of error vs total votes cast
Error magnitude
0
5
10
15
20
25
2e+05
5e+05
1e+06
2e+06
5e+06
1e+07
Total votes


EV: size of error vs total votes cast
Error magnitude
0
5
10
15
2e+05
5e+05
1e+06
2e+06
5e+06
1e+07
Total votes


Pollster: size of error vs total votes cast
Error magnitude
0
5
10
15
2e+05
5e+05
1e+06
2e+06
5e+06
1e+07
Total votes

So if the frequency of the disease is less than about 56 per million, the old test is better, on average.

For patients, no test is preferable to the new test when

$$\begin{aligned} 10^7 p &< 50 + (10^5 - 50)p \\ p &< 5.1 \times 10^{-6} \end{aligned}$$

However, since $5.1 \times 10^{-6} < 5.6 \times 10^{-5}$, under these circumstances pateints prefer the old test to the new test anyway, and the old test always beats no test, so no test is never preferred.

To sum up: patients prefer the new test, unless the probability of having the disease $p < 5.6 \times 10^{-5}$.

(c) *For an insurance company, the cost of follow-up testing for someone who does not have X but is falsely diagnosed with it is \$5,000, and the cost of treating someone who really does have X is \$100,000. Untreated X costs the insurance company \$250 in pain-killers. Find the expected cost to the insurance company of offering the new test, the old test, and no test at all, as functions of p.*

ANSWER:

$$\begin{aligned} \mathbf{E}\left[\text{no test}\right] &= 250p \\ \mathbf{E}\left[\text{old test}\right] &= 250p \times 0.1 + 10^5 \times p \times 0.9 \\ &= 90025p \\ \mathbf{E}\left[\text{new test}\right] &= 5 \times 10^3 \times (1-p) \times 0.01 + 250p \times 0.01 + 10^5 p \times 0.99 \\ &= 50 - 50p + 2.5p + 99000p \\ &= 50 + 99052.5p \end{aligned}$$

(d) *For what values of p do patients and insurance companies agree on which test to take?*

ANSWER: The insurance company would always rather give no test than the old test (since $99025 > 250$). So the only question is whether the company will ever prefer the new test to no test.

$$\begin{aligned} 50 + 99052.5p &< 250p \\ 50 + 98802.5p &< 0 \\ p &< -5 \times 10^{-4} \end{aligned}$$

Since $p > 0$, however, this means that the insurance company would always prefer giving no test to giving the new test.

Since the insurance company would always rather give no test, there is no value of p for which the insurer and the patients agree.