HW #8

8.6, 8.12, 8.19, 8.20, 8.25

Hayden Atchley

2022-11-25

8.6

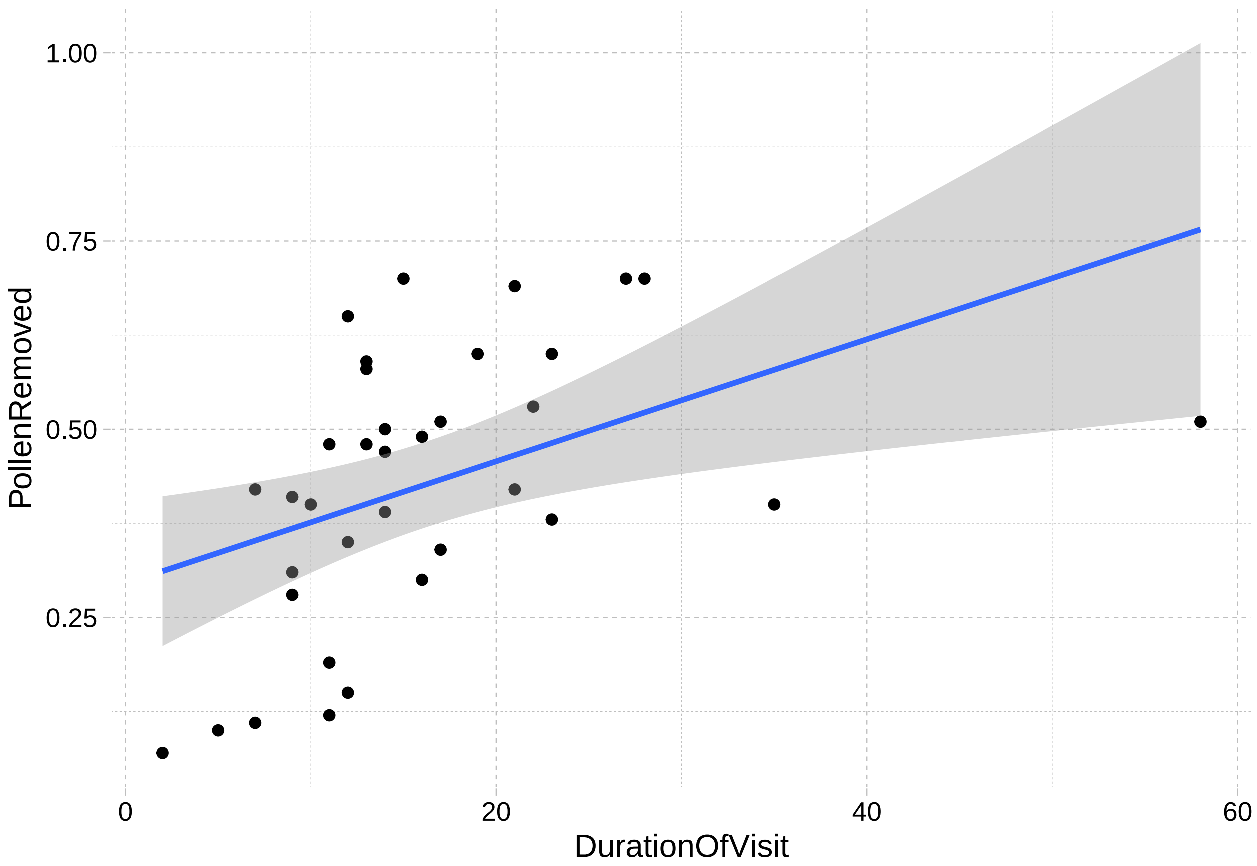
Using the volume as a group designator in an ANOVA test would only allow interpretation of results between these specific groups, whereas using volume as an explanatory variable in a regression model would allow for interpolation. Since in reality volume is a continuous, ordered scale rather than an unordered categorical designation, it is reasonable to use regression on this data.

8.12

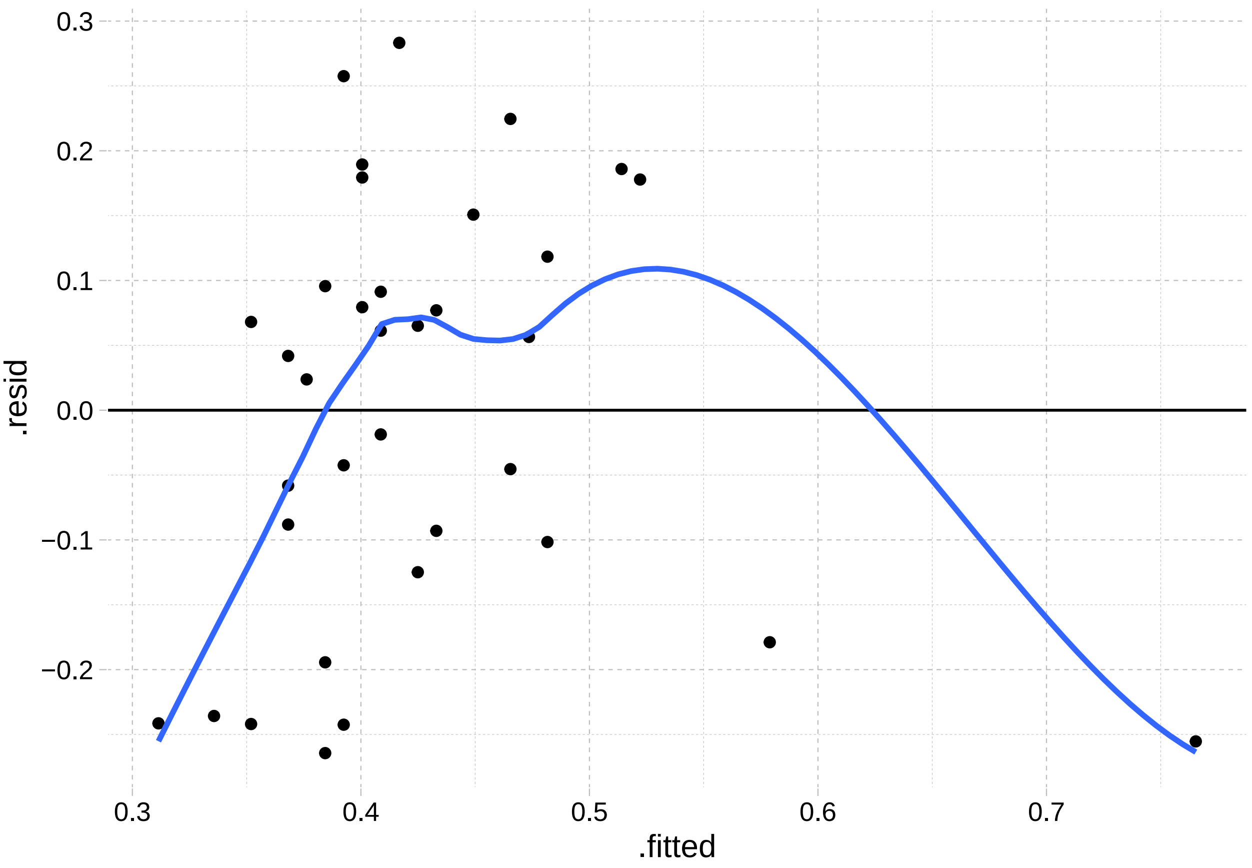
If there are no replicate responses, then the means at each input value would be equal to the single response, and there would be 0 degrees of freedom in the seperate-means model.

8.19

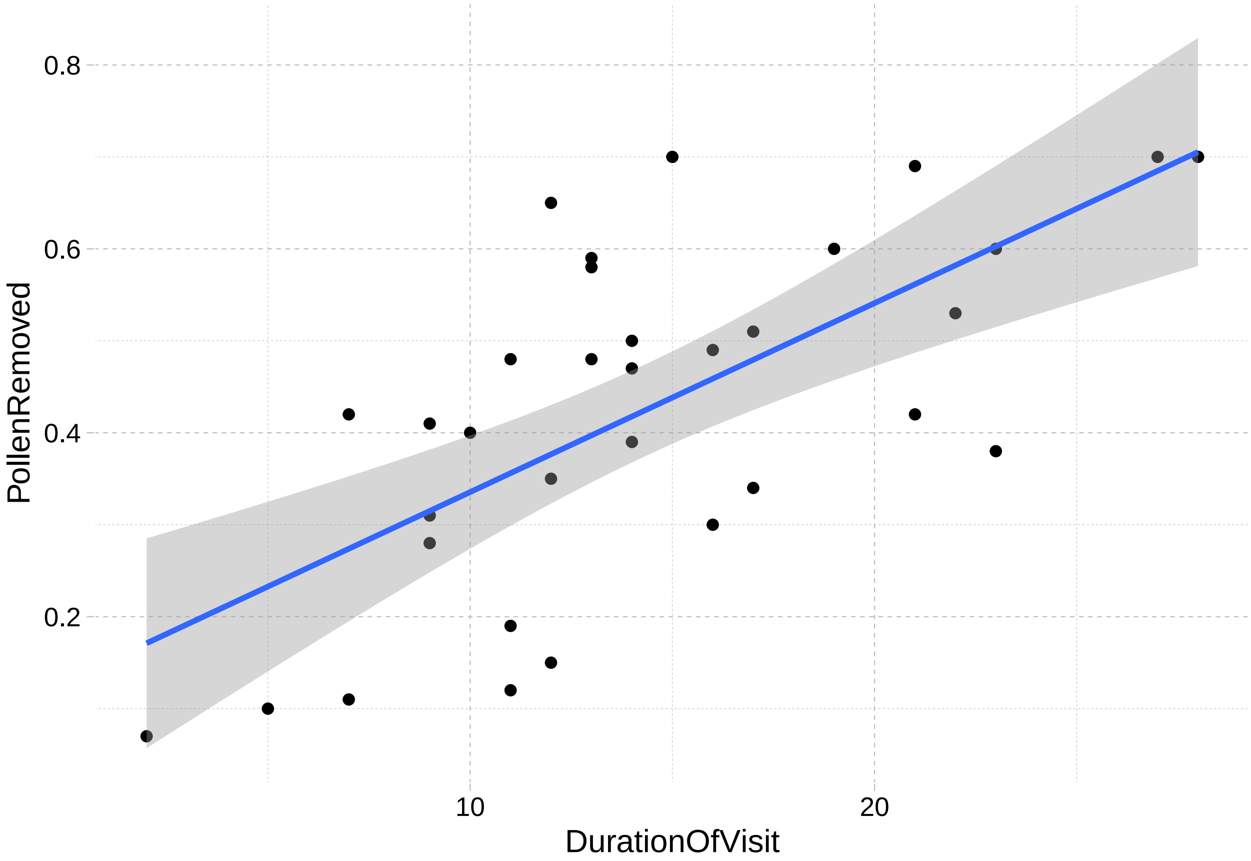
A plot of the pollen data:



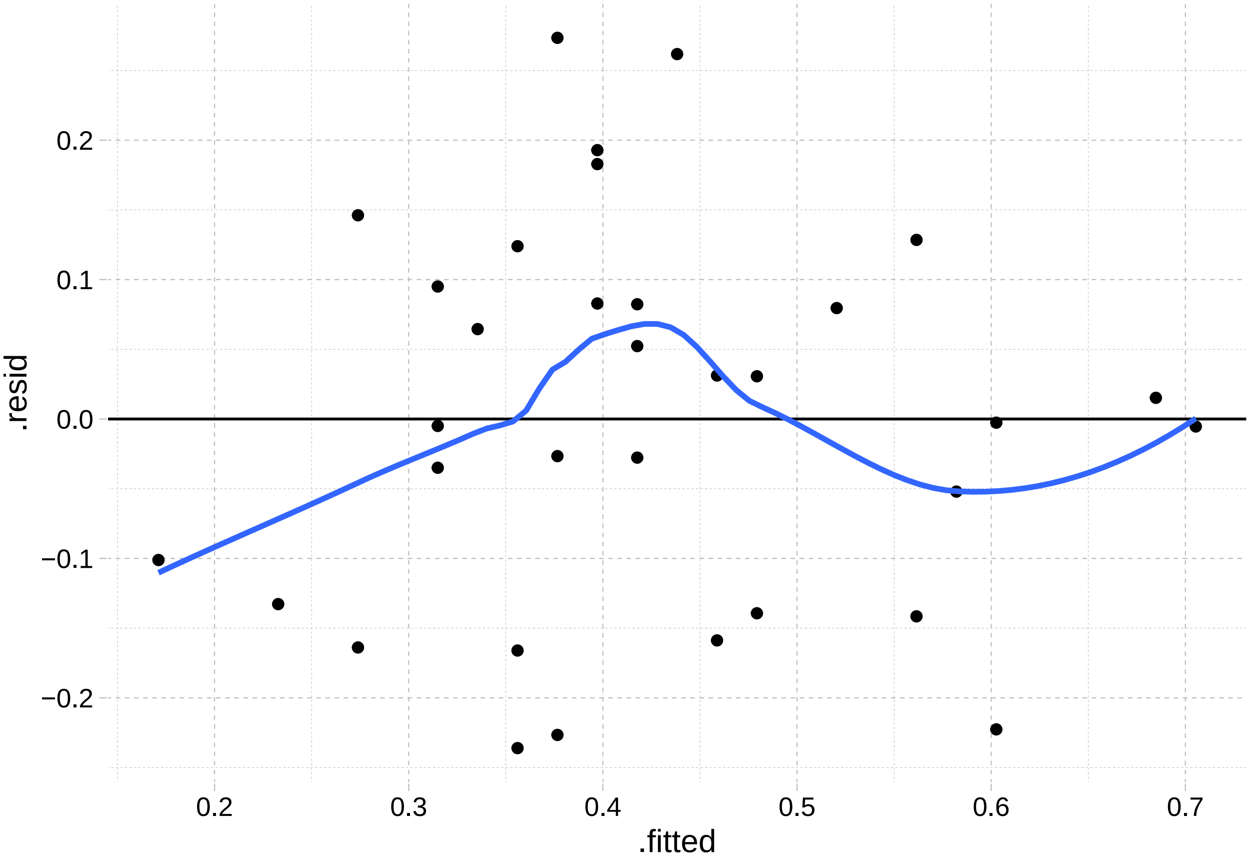
It’s clear from this plot that there are significant outliers, but to be sure we look at the residual plot:



This seems an obvious candidate for quadratic regression, and no log transformation of X or Y is likely to help. We fit the model again, removing times greater than 31 seconds:



And the residual plot:

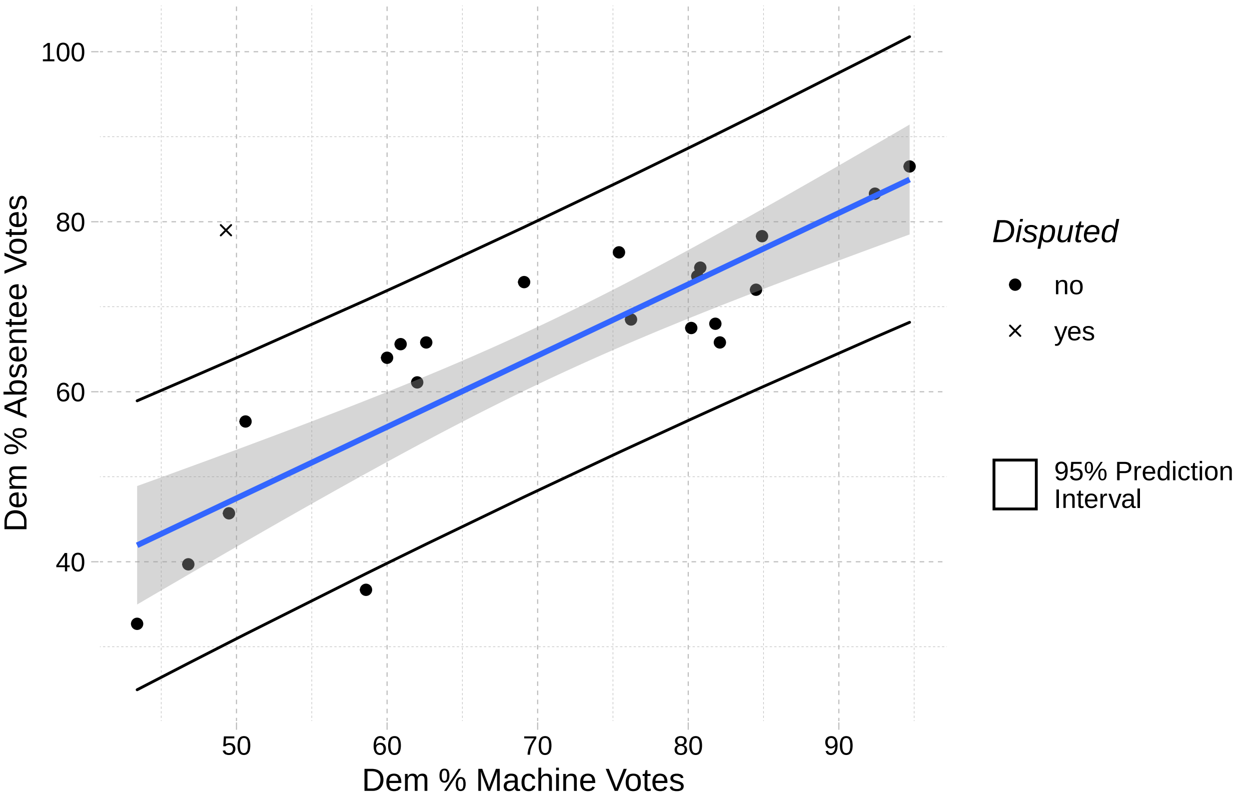


This fit looks significantly better, though a quadratic regression may still be an even better fit. A summary of the above model:

term	estimate	std.error	statistic	p.value
(Intercept)	0.130	0.0633	2.05	0.049
DurationOfVisit	0.021	0.0041	5.04	0.000

8.20

a



NOTE: The fit line and prediction interval on this model use only the non−disputed data points

The disputed election is well outside the 95% prediction interval of this model, though it is worth noting that it is not the only data point outside this range.

Looking at the fit and prediction of the specific disputed election gives us:

DemPctOfMachineVotes	DemPctOfAbsenteeVotes	fit	se.fit	df	residual.scale
49.3	79	46.9	2.79	19	7.41

which gives a *t*-score for the disputed result of $\frac{79.0-46.9}{2.79} = 11.5$. With 19 degrees of freedom, this gives a *p*-value of 2.65×10^{-10}, which is extremely low. But this is a cherry-picked data point, so we use the Bonferroni adjustment ($I = 22$ because there are 22 total residuals in the dataset we could analyze):

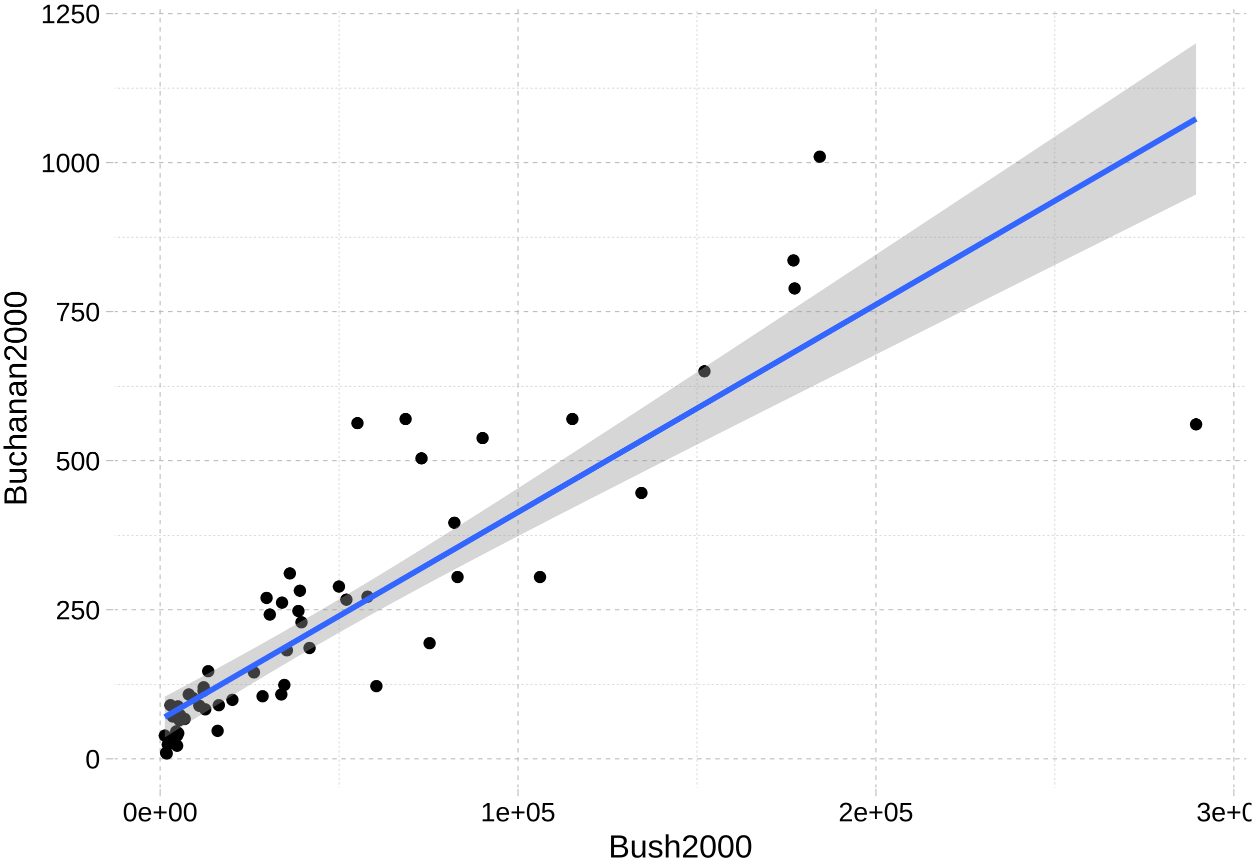
$$\begin{aligned}
k &= \frac{I \times (I-1)}{2} = \frac{22 \times 21}{2} = 231 \\
conf &= 0.95 = 1 - \alpha \implies \alpha = 0.05 \\
t\text{-multiplier} &= t_{d.f.}(1 - \alpha/2k) \qquad \text{(Bonferroni)} \\
&= t_{19}(1 - \frac{0.05}{2 \times 231}) = t_{19}(0.999892) \\
&= 4.55
\end{aligned}$$

We multiply this by the standard error to get a 95% confidence interval half-width of $2.79 \times 4.55 = 12.7$.

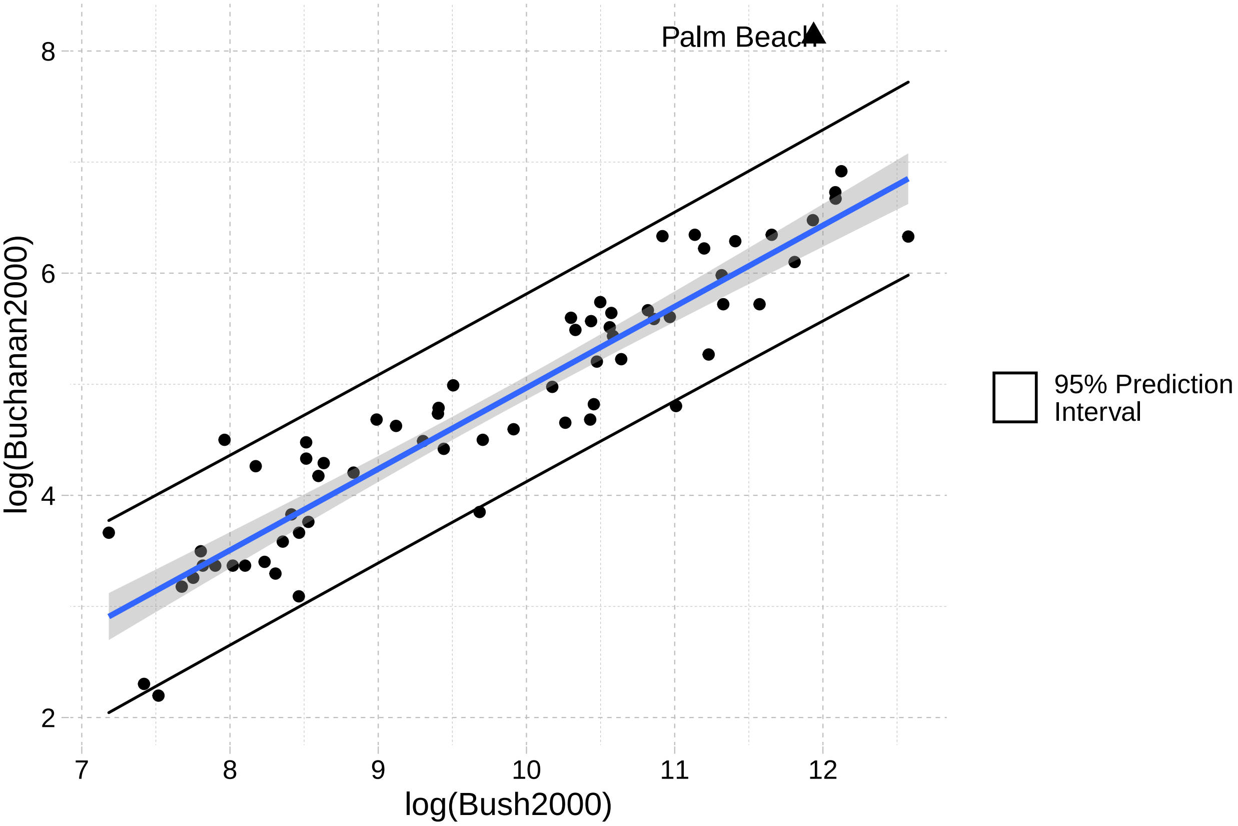
8.25

Because we are testing the results from Palm Beach against the rest, we will run our linear models excluding the data from Palm Beach.

Plotting this data subset gives:



The data points spread out as they increase in both x and y, so we plot the data again with both axes logged, and add the data from Palm Beach:



NOTE: The model excludes Palm Beach data

The model's upper 95% prediction value for the log(Bush2000) value that we see in Palm Beach is 7.24. The actual value of the Palm Beach data is 8.13. Because of this (and taking into account the log transformation), if the assumption that excess Buchannan votes were intended to be Gore votes, we can say with 95% confidence that at least $e^{8.13} - e^{7.24} = 2000$ Buchannan votes in Palm Beach were intended to be Gore votes.