STAT 223

Laboratory Activity #4 by Akbota Serikkyzy

Introduction

We will be considering how the dem_vote_margine is affected by a number of variables, such as unemployment, county_type, phd and avr_salary.

dem_vote_margin: The two-way margin the Democratic candidate received over the Republican candidate in a recent election
unemployment: The unemployment percentage in the county on the day of the election
county_type: The type of county as "rural", "suburban" and "urban"
phd: The percentage of the county with a PhD or other terminal degree
avg_salary: the average salary of county residents in thousands of dollars

Transformation

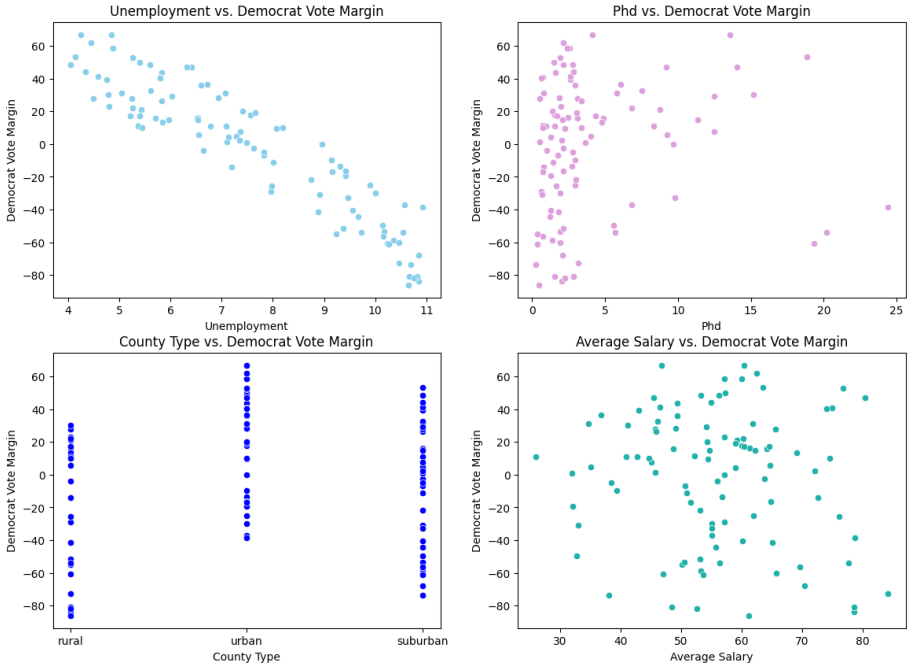

Figure 1: Scatter Plots

Looking at table 1, we can say that unemployment and Democrat vote margin have a strong negative linear relationship. This means that the higher the unemployment rate, the fewer people vote Democrat. When it comes to the phd percentage and Democrat vote margin, there is no clear linear relationship, and the data is heavily right-skewed, so trying log or polynomial transformations could be helpful. For county type vs. Democrat Vote Margin, there is clearly no linear relationship since we only have three categories for county type. From the scatter plot, we can observe that urban citizens tend to vote Democrat, rural residents lean more non-Democrat, and suburban areas fall somewhere in between. Since this is a categorical variable, no transformation is needed. When it comes to average salary and Democrat Vote Margin, the data is quite spread out, making it difficult to detect any clear pattern. In this case, it could be beneficial to try log and polynomial transformations to see if they improve the model's performance.

Multiple Linear Regression - Model 1

Variable	Coefficient	Standard Error	p-value
Intercept	107.9502	2.708	0.000
county_type[T.suburban]	16.1259	1.116	0.000
county_type[T.urban]	37.5384	1.119	0.000
unemployment	-17.4326	0.212	0.000
phd	0.5948	0.092	0.000
avg_salary	0.0006	0.038	0.988

Table 1: Regression Results for Democratic Vote Margin

Statistic	Value
R-squared	0.989
Adjusted R-squared	0.988
F-statistic	1689
Prob (F-statistic)	2.28e-90
AIC	583.8

Table 2: Model Summary Statistics for Regression

$$\begin{aligned}\text{dem vote margin} = {} & 107.9502 + 16.1259 \times \text{suburban} + 37.5384 \times \text{urban} - 17.4326 \times \text{unemployment} \\ & + 0.5948 \times \text{phd} + 0.0006 \times \text{avg salary}\end{aligned}$$

Looking at the table 1, we can see that most of the predictors are statistically significant, as their p-values are close to 0. However, average salary is statistically insignificant, with a p-value of 0.988, suggesting it does not have a meaningful effect on the Democrat vote margin in this model. The predictors county_type[T.suburban] and county_type[T.urban] are highly positively associated with the Democrat vote margin, indicating that counties categorized as suburban and urban tend to

have higher Democratic support compared to rural ones.

Both PhD attainment and average salary show a positive relationship with the Democrat vote margin, though the effects are small. Interestingly, unemployment is the only variable in the model with a negative coefficient, implying that higher unemployment is associated with a decrease in Democratic vote margin. This might suggest that economic dissatisfaction could be reason for reduced support for Democratic candidates in certain counties.

Turning to the model summary statistics in table 2, we observe a high R-squared value of 0.989. This indicates that 98.9% of the variance in the Democrat vote margin is explained by the predictors in the model, which reflects a very strong fit. The p-value of the overall model is extremely close to 0, meaning the model is statistically significant overall. The AIC value is 583.8.

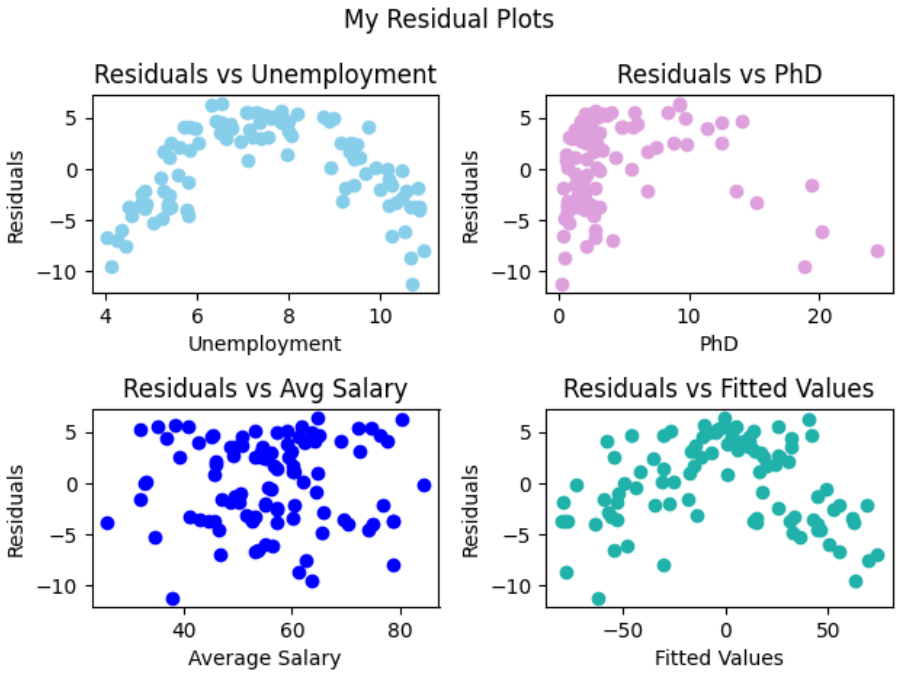

Figure 2: Resudual Plots

Looking at the figure 2, we can see that unemployment does not have a linear relationship with the Democrat vote margin, as the residuals form a reversed U-shape. This suggests a non-linear pattern, where the model might be under-predicting at the low and high ends of unemployment and over-predicting in the middle. Similarly, PhD attainment shows signs of heteroscedasticity, meaning the spread of residuals increases or decreases with the fitted values. While it is not as clearly curved as unemployment, the pattern still hints at possible nonlinearity.

When it comes to the average salary, the residuals are widely spread out but relatively evenly distributed, suggesting a more consistent and possibly linear relationship between the average salary and the vote margin, although its coefficient was statistically insignificant in the regression results.

As for the residual vs. fitted values plot, the presence of a slight reversed U-shape in the residuals vs. fitted plot indicates that the model may not be capturing some non-linear trends in the data. This kind of pattern suggests that that the model is systematically over-predicting or under-predicting in certain ranges of the outcome, which could be improved by transforming variables.

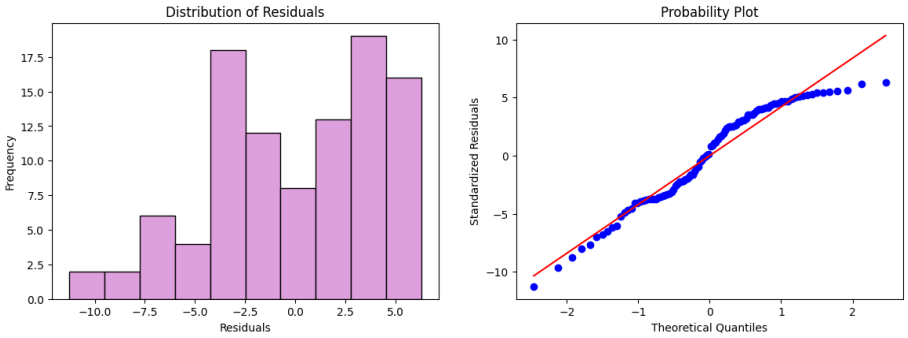

Figure 3: Resudual Plots

Looking at the figure 3, we can see that the residual plot appears to be left-skewed, meaning there are more negative residuals than positive ones. This suggests that the model tends to underestimate the actual values more often than it overestimates them. When we look at the Q-Q plot, we notice that the blue line does not consistently follow the red reference line. This deviation indicates that the residuals are not perfectly normally distributed, which might be a concern for linear regression.

Quadratic Term - Model 2

Variable	Coefficient	Standard Error	p-value
Intercept	48.4417	3.479	0.000
county_type[T.suburban]	15.4354	0.523	0.000
county_type[T.urban]	36.1080	0.529	0.000
unemployment	-1.0237	0.899	0.258
np.square(unemployment)	-1.0751	0.059	0.000
phd	0.7347	0.044	0.000
avg_salary	0.0317	0.018	0.078

Table 3: Regression Results for Democratic Vote Margin (Quadratic Model)

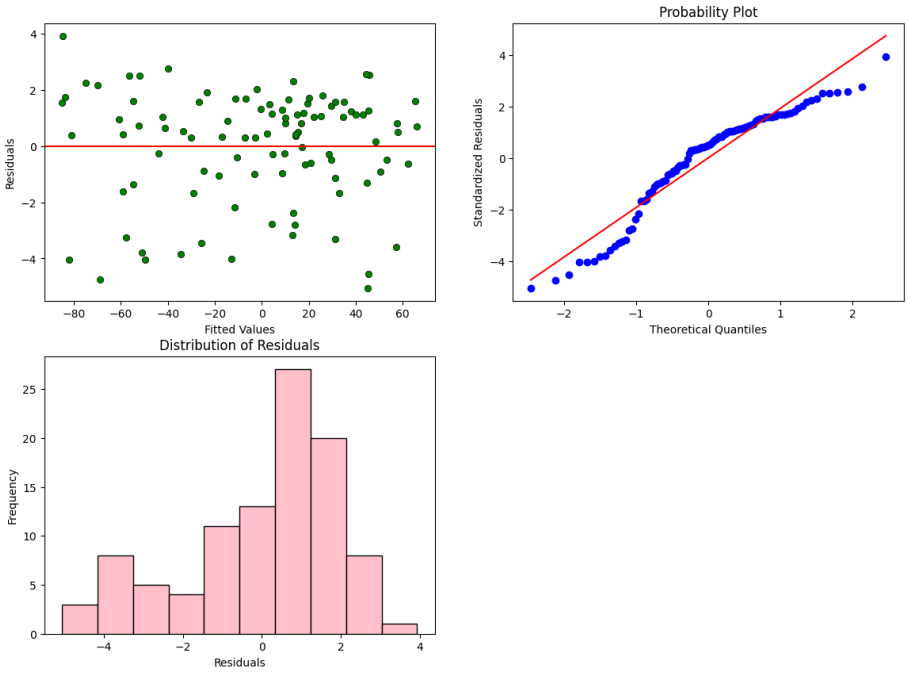

Figure 4: Resudual Plots

Statistic	Value
R-squared	0.998
Adjusted R-squared	0.997
F-statistic	6499
Prob (F-statistic)	1.16e-119
AIC	432.6

Table 4: Model Summary Statistics for Quadratic Regression

$$\begin{aligned}\text{dem vote margin} = 48.4417 + 15.4354 \times \text{suburban} + 36.1080 \times \text{urban} - 1.0237 \times \text{unemployment} \\ - 1.0751 \times \text{unemployment}^2 + 0.7347 \times \text{phd} + 0.0317 \times \text{avg salary}\end{aligned}$$

This is the model after adding a quadratic term for unemployment. As we can observe in table 3, the p-values slightly changed, with most predictors still close to 0, indicating strong statistical significance. However, both unemployment and average salary are now statistically insignificant. Although we only added a quadratic term for unemployment, it appears to have also impacted the significance of average salary. This could be due to multicollinearity. Adding a new variable that is correlated with the original unemployment variable could have changed the variance shared with other predictors like average salary, thus affecting their significance. Additionally, the coefficient for unemployment used to be around -17.4326, but after including the quadratic term, it dropped

to around -1. This shift is expected when adding a squared term, as the model now accounts for a curved relationship. In this case, the linear and quadratic terms work together to capture that non-linear pattern, so the individual linear coefficient naturally becomes smaller.

In general, most predictors remain statistically significant. Looking at the model summary in table 4, we see that R-squared increased from 0.989 to 0.998, indicating an even better fit. This means that 99.8% of the variation in the margin of democrat votes is now explained by the predictors in the model. The p-value remains extremely close to 0, suggesting that the overall model is statistically significant. Additionally, the AIC dropped from 583.8 to 432.6, which means this model balances complexity and explanatory power better.

Looking at figure 4, the Residuals vs Fitted Values plot shows that the residuals are spread fairly randomly around the horizontal line at zero without any patterns. This suggests that the model now better satisfies the assumption of linearity after adding the quadratic term for unemployment.

When it comes to the Q-Q plot, we can see that the blue points making waves around the red reference line. This suggests that the residuals are not perfectly normally distributed, which means we should try another transformation.

Looking at the histogram of residuals, it appears slightly left-skewed. However, the majority of residuals are clustered between 0 and 2, meaning that while the distribution has a longer left tail, the model does not really underestimate the outcomes. The slight skewness suggests some mild deviation from normality but is not extremely severe.

Overall, the model seems to have improved in terms of linearity, but some minor concerns remain regarding the normality of residuals.

Log Transformation - Model 3

Variable	Coefficient	Standard Error	p-value
Intercept	50.6076	1.194	0.000
county_type[T.suburban]	15.4736	0.181	0.000
county_type[T.urban]	35.7428	0.184	0.000
unemployment	-1.9330	0.309	0.000
np.square(unemployment)	-1.0029	0.020	0.000
np.log(phd)	4.0261	0.073	0.000
avg_salary	0.0260	0.006	0.000

Table 5: Regression Results for Democratic Vote Margin (Log-Transformed Model)

$$\begin{aligned}\text{dem vote margin} = 50.6076 + 15.4736 \times \text{suburban} + 35.7428 \times \text{urban} - 1.9330 \times \text{unemployment} \\ - 1.0029 \times \text{unemployment}^2 + 4.0261 \times \log(\text{phd}) + 0.0260 \times \text{avg salary}\end{aligned}$$

Statistic	Value
R-squared	1.000
Adjusted R-squared	1.000
F-statistic	54120
Prob (F-statistic)	2.01e-162
AIC	220.8

Table 6: Model Summary Statistics for Log-Transformed Regression

This is the model after applying a log transformation to the PhD variable. Looking at table 5, we see that all predictors are now statistically significant, with p-values very close to 0. Additionally, the coefficient for PhD changed notably. It used to be between 0 and 1 in previous models, but after the log transformation, it increased to 4.0261. This change is natural since log transforming a variable changes the scale of its effect.

Looking at the table 6, R-squared became 1.000, which means that our predictors are responsible for 100% of the variance in the Democratic vote margin. Adjusted R-squared supports this idea, also being 1.000. The p-value for the model is again super close to 0, meaning our model is statistically significant. AIC dropped to 138, in previous models, it used to be 583.8 and 432.6. So, it is a big drop, meaning our model now works better in terms of fit and simplicity.

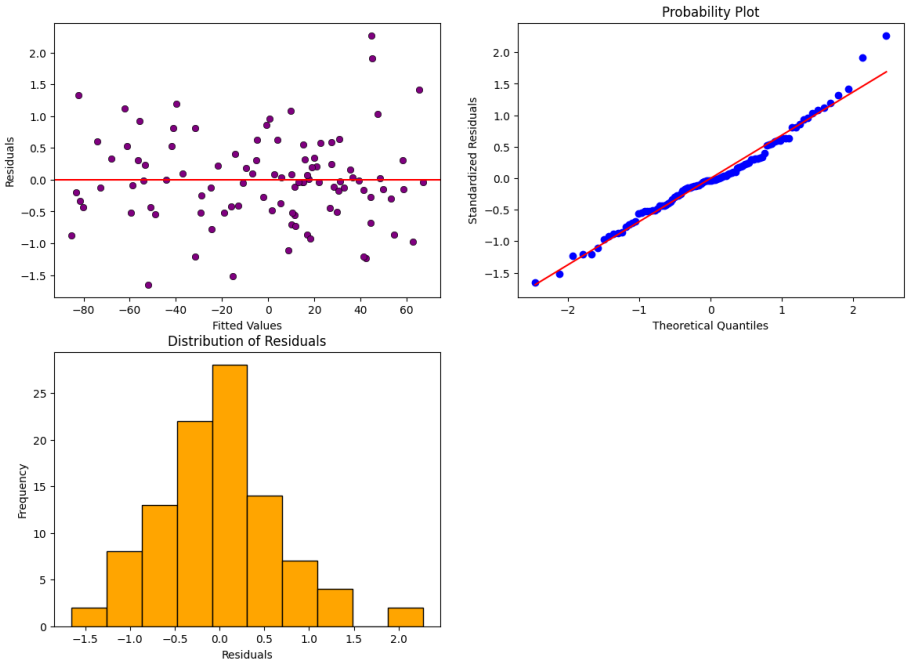

Figure 5: Resudual Plots

Looking at figure 5, the residuals vs fitted values plot supports the idea of linearity since the residuals are spread out randomly without any clear pattern. When it comes to the Q-Q plot, the residuals now appear to be normally distributed because the blue points mostly follow the red line, with minor deviations at the tails. The histogram is roughly symmetric, confirming the observations from the Q-Q plot. Overall, the residual plots suggest that the model assumptions are fairly well satisfied, though there are slight outliers in the tails.

Interaction Terms - Model 4

Variable	Coefficient	Standard Error	p-value
Intercept	50.0818	0.787	0.000
county_type[T.suburban]	15.9627	0.152	0.000
county_type[T.urban]	34.8788	0.176	0.000
unemployment	-1.7489	0.203	0.000
np.square(unemployment)	-1.0174	0.013	0.000
np.log(phd)	4.0772	0.083	0.000
county_type[T.suburban]:np.log(phd)	-0.5825	0.110	0.000
county_type[T.urban]:np.log(phd)	0.7774	0.127	0.000
avg_salary	0.0258	0.004	0.000

Table 7: Regression Results for Democratic Vote Margin with Interaction Terms

Statistic	Value
R-squared	1.000
Adjusted R-squared	1.000
F-statistic	94640
Prob (F-statistic)	7.65e-175
AIC	138.0

Table 8: Model Summary Statistics for Interaction Term Regression

$$\begin{aligned}\text{dem vote margin} = {} & 50.0818 + 15.9627 \times \text{suburban} + 34.8788 \times \text{urban} \\ & - 1.7489 \times \text{unemployment} - 1.0174 \times \text{unemployment}^2 \\ & + 4.0772 \times \log(\text{phd}) - 0.5825 \times (\text{suburban} \times \log(\text{phd})) \\ & + 0.7774 \times (\text{urban} \times \log(\text{phd})) + 0.0258 \times \text{avg salary}\end{aligned}$$

This model includes interaction terms. After testing various combinations of predictors, the model that performed best featured the interaction between county_type and log(phd). As shown in table 7, all predictors are statistically significant, including the interaction terms between phd and county_type. This suggests that the effect of educational background on Democratic vote margin

varies depending on the type of county.

Looking at table 8, the R-squared remains 1.000, indicating that the model explains 100% of the variation in Democratic vote margin. The p-value is close to zero, suggesting the overall model is statistically significant. Also, the AIC remains at 138.0, consistent with Model 3, indicating that including interaction terms improved model interpretability without affecting model complexity.

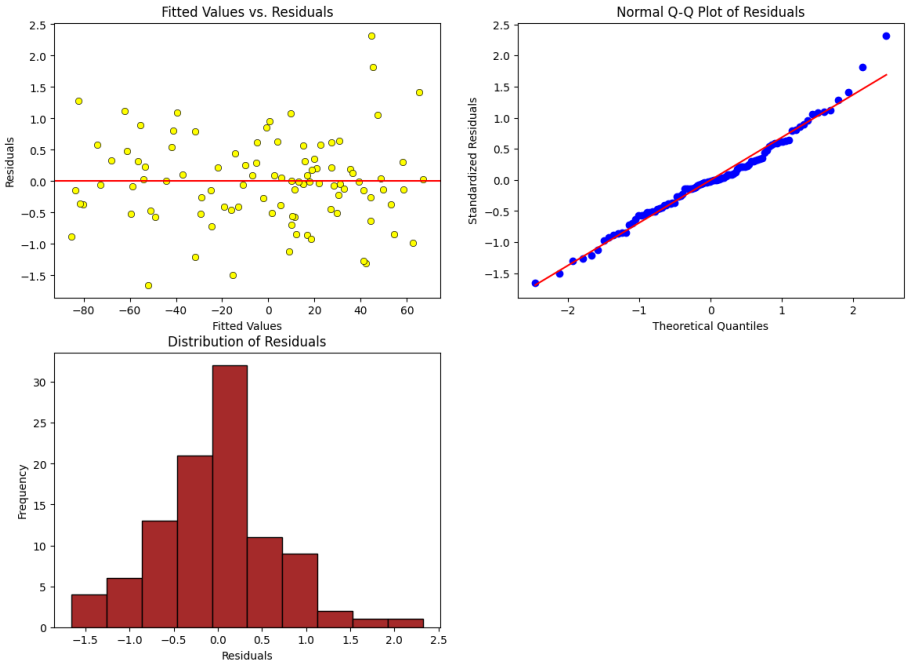

Figure 6: Resudual Plots

Looking at figure 6, the residuals vs fitted values plot supports linearity since the residuals are spread out randomly without any clear pattern. When it comes to the Q-Q plot, the residuals now appear to be normally distributed because the blue points mostly follow the red line, with minor deviations at the tails. The histogram look approximately normal, but with a slight positive skew. This suggests the model is performing well, though the assumption of perfect normality is violated a bit. Overall, the residual plots suggest that the model satisfies the assumptions of linearity and approximate normality, showing only a slight positive skew in the histogram.

Conclusion

In conclusion, throughout the process we saw improvements. At first, we noticed signs of non-linearity, so we worked on improving the model in terms of normality and linearity. The best performing models were the last two models which are model 3 with a log transformation of PhD

and model 4 with interaction terms, including between county type and PhD.

The interaction terms in model 4 added some complexity, but its performance was very close to model 3. However, the histogram for model 4 showed a bit more abnormality compared to the more normal residuals in model 3. Therefore, in terms of simplicity and assumptions, model 3 performed the best. However, if we are also interested in the relationship between PhD and county type, then model 4 is a good choice too.

One thing to note is that the R-squared is 1.000 in our model 3 and model 4, which might mean our model is overfitting, and we should take this into consideration in the future steps.