Regression Review

Spring 2024

RESEARCH QUESTIONS

1. Is there an effect of news exposure on political knowledge?

2. Is there an effect of news exposure on political knowledge after controlling for demographic and political covariates?

3. Does education level *moderate* the effect of news exposure on political knowledge after controlling for demographic and political covariates?

DATA

The data in *pew.csv* come from a telephone survey conducted by The Pew Research Center for The People & The Press in February 2007. The data represent a probability sample of 1,502 adults in the U.S.

```
# A tibble: 1,502 × 10
      id knowledge  news   age  educ  male ideology party    state_blue engagement
   <dbl>     <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl> <chr>         <dbl>      <dbl>
 1     1        50    63  44.4    14     0     39.3 Democrat          0       79.5
 2     5        79    54  59.5    15     1     69.5 Democrat          0       80.8
 3     6        31    65  43.8    12     0     18   Democrat          0       99.6
 4     8        93    50  29.6    16     1     29.7 Democrat          0       88.6
 5     9        23    12  45.1    12     0     52.5 Democrat          0       82.4
 6    11        23    43  47.5     9     0     62   Democrat          0       67
 7    13        44    34  18.3    10     1     28.3 Democrat          0       69.4
 8    14        97    56  79.1    17     1     30.1 Democrat          0       78.2
 9    16        54    81  32.5    13     0     31.7 Democrat          0       97.8
10    18        58    66  30.8    16     0     11   Democrat          0       78.3
# ℹ 1,492 more rows
# ℹ Use `print(n = ...)` to see more rows
```

CATEGORIZE ATTRIBUTES

Outcome
Political knowledge (`knowledge`)

Focal Predictor(s)
News exposure (`news`)

Covariate(s)

Demographic
- Age (`age`)
- Education level (`educ`)
- Sex (`male`)

Political
- Political ideology (`ideology`)
- Political affiliation (`party`)
- Blue/red state (`state_blue`)
- Political engagement (`engagement`)

General analytic strategy

Explore Attributes

Examine attributes (plots, numerical summaries)

1. Outcome
2. Focal predictor
3. Covariates

Explore Relationships

Examine relationship between focal predictor/covariates and outcome

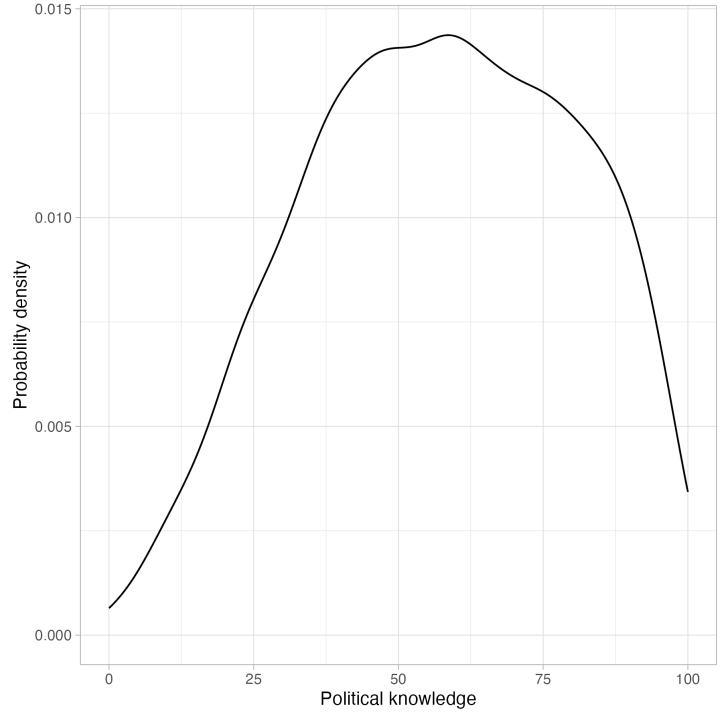
1. Scatterplot
2. Correlations

Modeling

Fit models

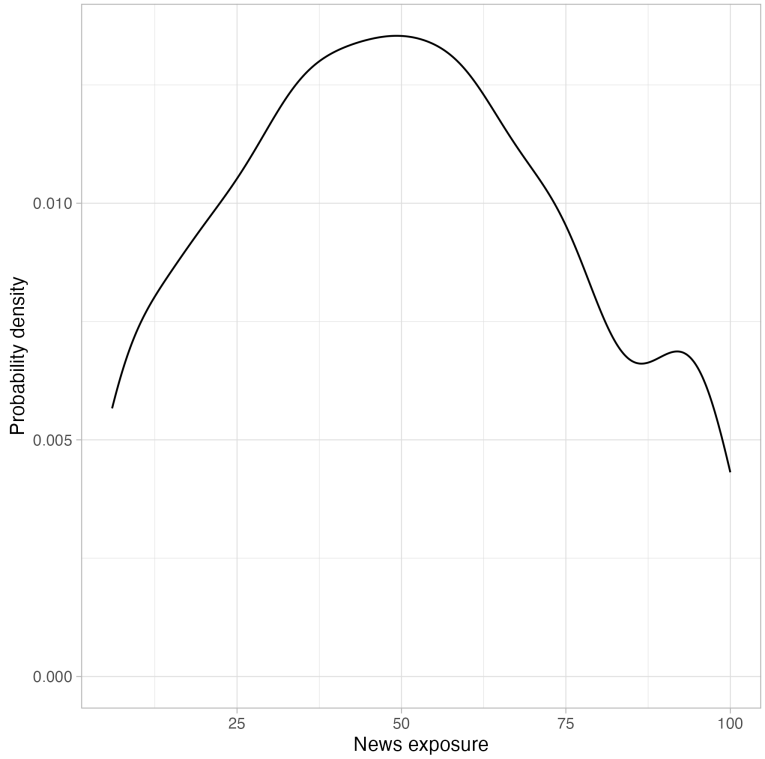
1. Examine residuals
2. Model-level output (`glance()`)
3. Coefficient-level output (`tidy()`)

OUTCOME



M	SD
57.1	22.9

FOCAL PREDICTOR

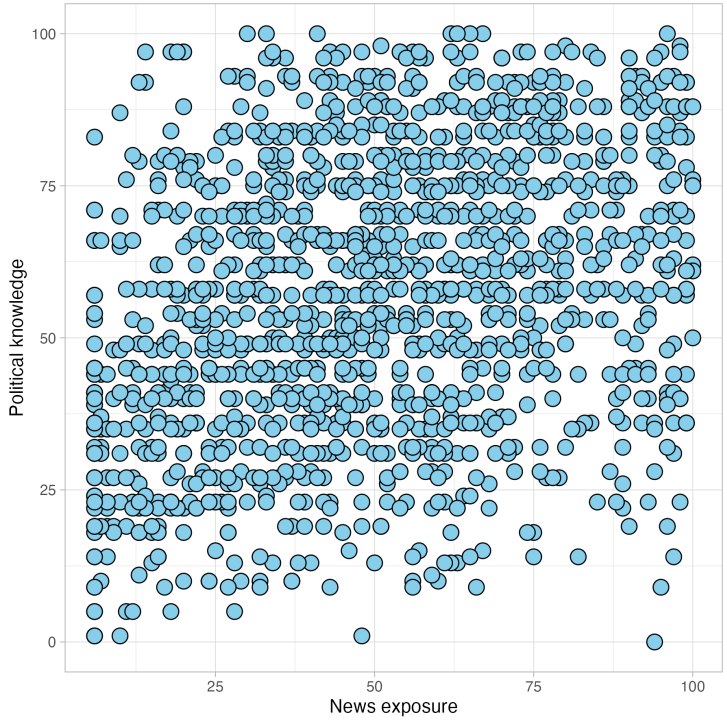


M	SD
50.3	25.3

COVARIATES

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
1 age	0	1	50.9	17.9	18	37.6	50.2	64.5	95.3	
2 educ	0	1	14.0	2.41	8	12	14	16	18	
3 male	0	1	0.491	0.500	0	0	0	1	1	
4 ideology	0	1	46.7	20.0	0	31.7	46.4	58.1	100	
5 state_blue	0	1	0.429	0.495	0	0	0	1	1	
6 engagement	0	1	73.4	16.8	7.5	64	76.8	85.4	100	
7 democrat	0	1	0.362	0.481	0	0	0	1	1	
8 republican	0	1	0.479	0.500	0	0	0	1	1	

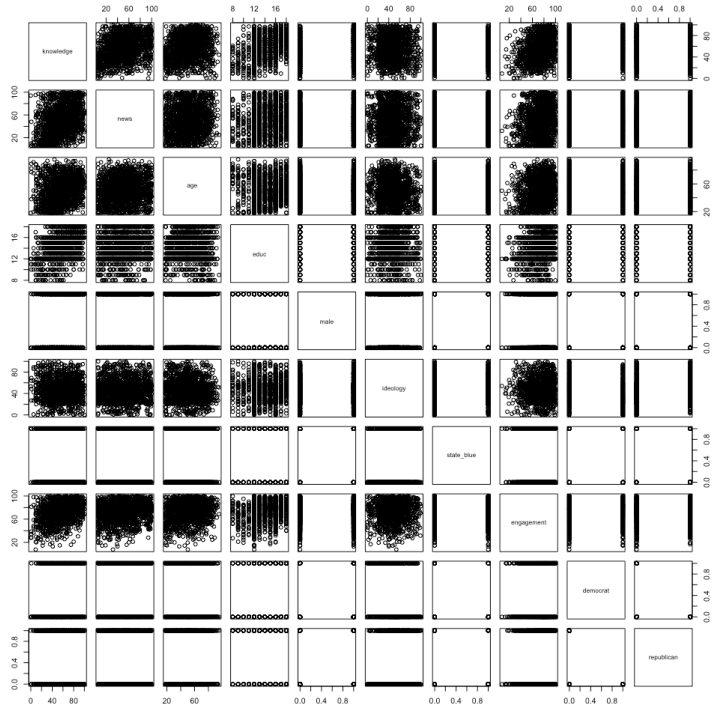
RELATIONSHIP BETWEEN FOCAL PREDICTOR AND OUTCOME



r = 0.322

RELATIONSHIP BETWEEN ALL NUMERIC ATTRIBUTES

This is called a scatterplot matrix. It shows the scatterplot for every pair of attributes.



CORRELATIONS BETWEEN ALL NUMERIC ATTRIBUTES

	term	knowledge	news	age	educ	male	ideology	state_blue	engagement	democrat	republican
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	knowledge	NA	0.322	0.217	0.450	0.239	-0.0212	0.0720	0.332	0.0826	0.0578
2	news	0.322	NA	0.119	0.144	0.0282	0.0275	-0.00315	0.168	-0.0379	0.127
3	age	0.217	0.119	NA	0.00646	-0.0223	-0.0542	0.104	0.163	0.00647	0.0354
4	educ	0.450	0.144	0.00646	NA	0.0347	0.0654	0.0733	0.247	0.0449	0.0116
5	male	0.239	0.0282	-0.0223	0.0347	NA	-0.0625	-0.0148	-0.0535	0.0542	-0.0794
6	ideology	-0.0212	0.0275	-0.0542	0.0654	-0.0625	NA	0.119	0.0143	-0.329	0.310
7	state_blue	0.0720	-0.00315	0.104	0.0733	-0.0148	0.119	NA	0.00405	-0.0817	0.0760
8	engagement	0.332	0.168	0.163	0.247	-0.0535	0.0143	0.00405	NA	0.0526	0.102
9	democrat	0.0826	-0.0379	0.00647	0.0449	0.0542	-0.329	-0.0817	0.0526	NA	-0.721
10	republican	0.0578	0.127	0.0354	0.0116	-0.0794	0.310	0.0760	0.102	-0.721	NA

RQ1: IS THERE AN EFFECT OF NEWS EXPOSURE ON POLITICAL KNOWLEDGE?

Statistical Evidence: Statistically relevant effect of news exposure in model that explains variation in political knowledge.

$$\text{Knowledge}_i = \beta_0 + \beta_1(\text{News Exposure}_i) + \epsilon_i \quad \text{where } \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$$

MODEL-LEVEL OUTPUT

RMSE

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.103	0.103	21.7	173.	1.72e-37	1	-6753.

AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<int>	<int>
13513.	13529.	707335.	1500	1502

Differences in the amount of news exposure explains 10.3% of the variation in political knowledge, which is more than we would expect because of chance; $F(1, 1500) = 173$, $p < .001$.

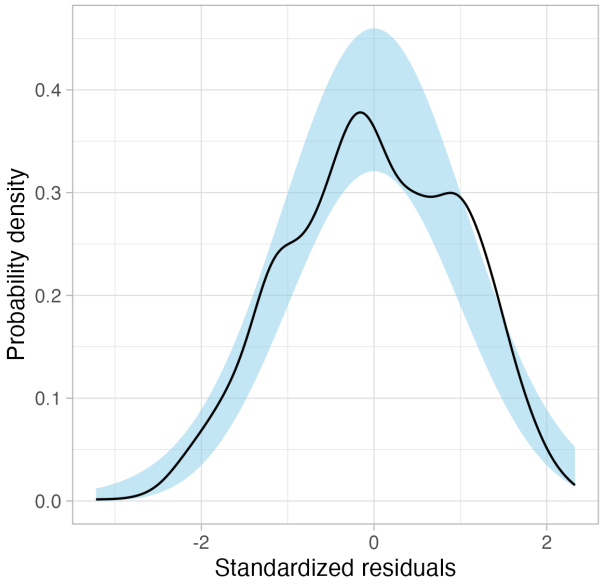
COEFFICIENT-LEVEL OUTPUT

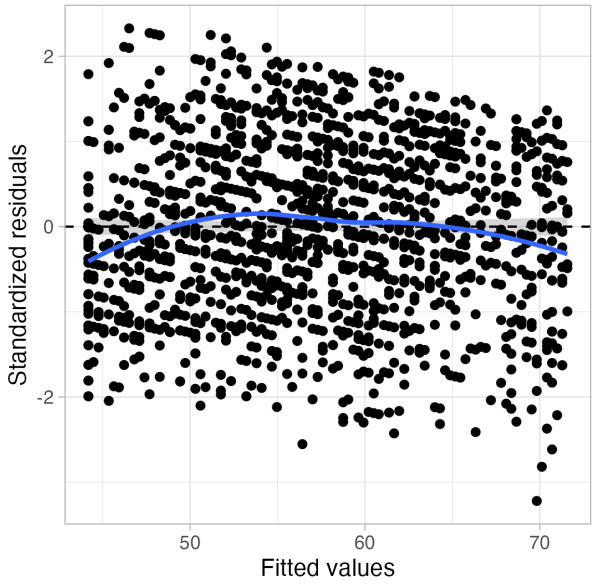
	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	42.4	1.25	34.0	9.46e-189
2	news	0.291	0.0221	13.2	1.72e- 37

Intercept: The political knowledge score for U.S. adults who have no news exposure is 42.4, on average.

Each 1-unit increase in news exposure is associated with a 0.29-unit increase in political knowledge, on average. The news exposure effect is more than we expect because of chance ($p < .001$)

RESIDUAL PLOTS FOR MODEL A





There seems to be some violation of the linearity assumption.

Also the normality assumption may be violated.

Don't forget about the independence assumption!!

RQ2: Is there an effect of news exposure on political knowledge after controlling for demographic and political covariates?

Statistical Evidence: Statistically relevant effect of news exposure in main-effects model that explains variation in political knowledge that also includes all the demographic and political covariates.

$$\text{Knowledge}_i = \beta_0 + \beta_1(\text{Age}_i) + \beta_2(\text{Education}_i) + \beta_3(\text{Male}_i) + \beta_4(\text{Engagement}_i) + \beta_5(\text{Ideology}_i) + \beta_6(\text{Democrat}_i) + \beta_7(\text{Republican}_i) + \beta_8(\text{Blue State}_i) + \beta_9(\text{News Exposure}_i) + \epsilon_i$$

where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$

MODEL-LEVEL OUTPUT

RMSE

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.398	0.394	17.8	109.	2.49e-157	9	-6455.

AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<int>	<int>
12932.	12990.	475274.	1492	1502

Differences in **all** of the predictors (i.e., in the model) explains 39.8% of the variation in political knowledge, which is more than we would expect because of chance; $F(9, 1492) = 109$, $p < .001$.

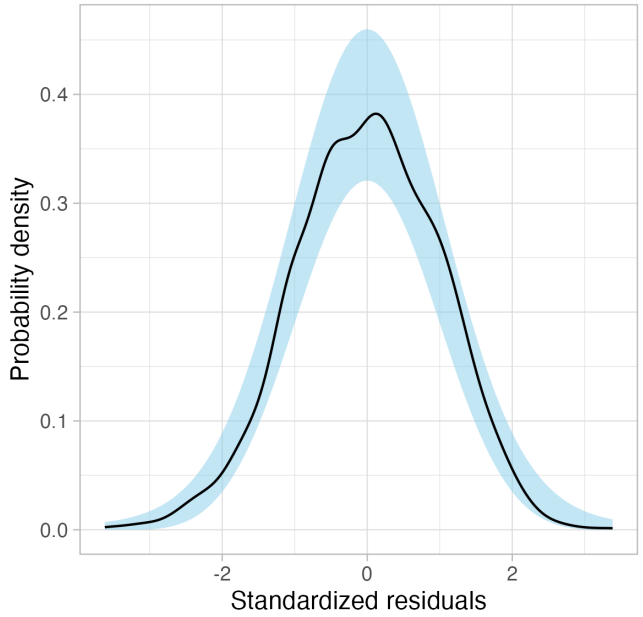
COEFFICIENT-LEVEL OUTPUT

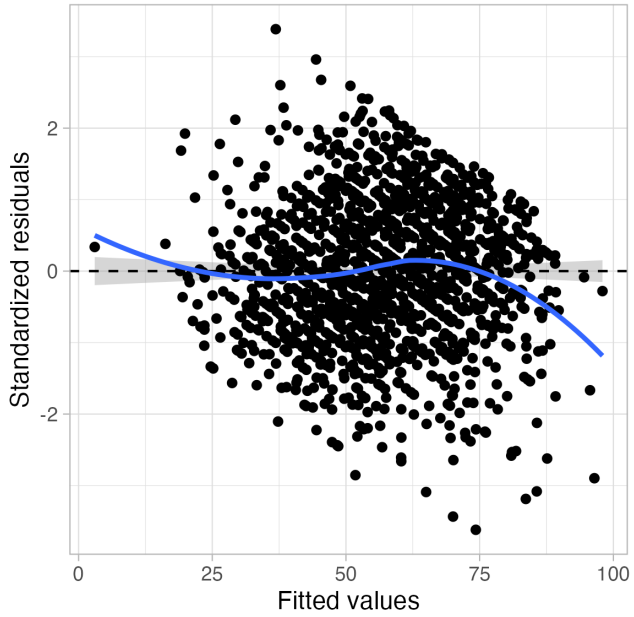
	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-37.3	3.48	-10.7	7.15e-26
2	age	0.199	0.0264	7.54	8.23e-14
3	educ	3.42	0.200	17.1	9.76e-60
4	male	10.8	0.928	11.6	7.20e-30
5	engagement	0.240	0.0294	8.16	7.28e-16
6	ideology	-0.0329	0.0248	-1.33	1.85e- 1
7	democrat	6.53	1.43	4.57	5.35e- 6
8	republican	5.85	1.38	4.25	2.30e- 5
9	state_blue	1.75	0.947	1.85	6.51e- 2
10	news	0.186	0.0188	9.88	2.47e-22

We generally will keep covariates in the model (and report them) regardless of statistical significance!

Each 1-unit increase in news exposure is associated with a 0.19-unit increase in political knowledge, on average, after controlling for the set of demographic and political covariates. The partial effect of news exposure is more than we expect because of chance ($p < .001$)

RESIDUAL PLOTS FOR MODEL B





There seems to be some violation of the linearity assumption.

Also the homoscedasticity assumption may be violated.

RQ3: DOES EDUCATION LEVEL MODERATE THE EFFECT OF NEWS EXPOSURE ON POLITICAL KNOWLEDGE AFTER CONTROLLING FOR DEMOGRAPHIC AND POLITICAL COVARIATES?

Statistical Evidence: Statistically relevant interaction effect of news exposure and education in interaction model that explains variation in political knowledge and includes all the demographic and political covariates.

$$\text{Knowledge}_i = \beta_0 + \beta_1(\text{Age}_i) + \beta_2(\text{Education}_i) + \beta_3(\text{Male}_i) + \beta_4(\text{Engagement}_i) + \beta_5(\text{Ideology}_i) +$$
$$\beta_6(\text{Democrat}_i) + \beta_7(\text{Republican}_i) + \beta_8(\text{Blue State}_i) + \beta_9(\text{News Exposure}_i) +$$
$$\beta_{10}(\text{News Exposure}_i)(\text{Education}_i) + \epsilon_i$$

where $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$

MODEL-LEVEL OUTPUT

RMSE

	r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.399	0.395	17.8	99.1	3.01e-157	10	-6453.

AIC	BIC	deviance	df.residual	nobs
<dbl>	<dbl>	<dbl>	<int>	<int>
12929.	12993.	473870.	1491	1502

Differences in **all** of the predictors (i.e., in the model) explains 39.9% of the variation in political knowledge, which is more than we would expect because of chance; $F(10, 1491) = 99.1$, $p < .001$.

COEFFICIENT-LEVEL OUTPUT

	term	estimate	std.error	statistic	p.value
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>
1	(Intercept)	-48.5	6.35	-7.63	4.02e-14
2	age	0.203	0.0265	7.69	2.74e-14
3	educ	4.21	0.427	9.86	2.92e-22
4	male	10.9	0.929	11.7	1.76e-30
5	engagement	0.239	0.0294	8.14	8.44e-16
6	ideology	-0.0309	0.0248	-1.25	2.12e- 1
7	democrat	6.51	1.43	4.56	5.54e- 6
8	republican	5.82	1.38	4.23	2.49e- 5
9	state_blue	1.75	0.946	1.85	6.41e- 2
10	news	0.407	0.107	3.81	1.45e- 4
11	educ:news	-0.0159	0.00757	-2.10	3.57e- 2

To better interpret the interaction effect, plot it.

The effect of news exposure on political knowledge differs depending on education level after controlling for the set of demographic and political covariates. This interaction effect differs from 0 more than we expect because of chance (p = .036).

PLOTTING THE INTERACTION EFFECT

- Place the effect of news exposure on the *x*-axis (it is the focal predictor)
- Place political knowledge (outcome) on *y*-axis
- Choose multiple levels (2–3) of education level
- Choose interpretable values for the other covariates (1 value per covariate)

- Education level = {12, 16}
- Age = 50
- Sex = 0 (female)
- Political ideology = 0 (Liberal)
- Democrat = 1
- Republican = 0
- Blue state = 1 (Lives in blue state)
- Political engagement = 50 (moderately engaged)

EDUCATION = 12 (HIGH SCHOOL EDUCATION)

$$\widehat{\text{Knowledge}}_i = -48.5 + 0.20(\text{Age}_i) + 4.21(\text{Education}_i) + 10.9(\text{Male}_i) + 0.24(\text{Engagement}_i) + -0.03(\text{Ideology}_i) + 6.51(\text{Democrat}_i) + 5.82(\text{Republican}_i) + 1.75(\text{Blue State}_i) + 0.41(\text{News Exposure}_i) - 0.02(\text{News Exposure}_i)(\text{Education}_i)$$

$$\widehat{\text{Knowledge}}_i = -48.5 + 0.20(50) + 4.21(12) + 10.9(0) + 0.24(50) - 0.03(0) + 6.51(1) + 5.82(0) + 1.75(1) + 0.41(\text{News Exposure}_i) - 0.02(\text{News Exposure}_i)(12)$$

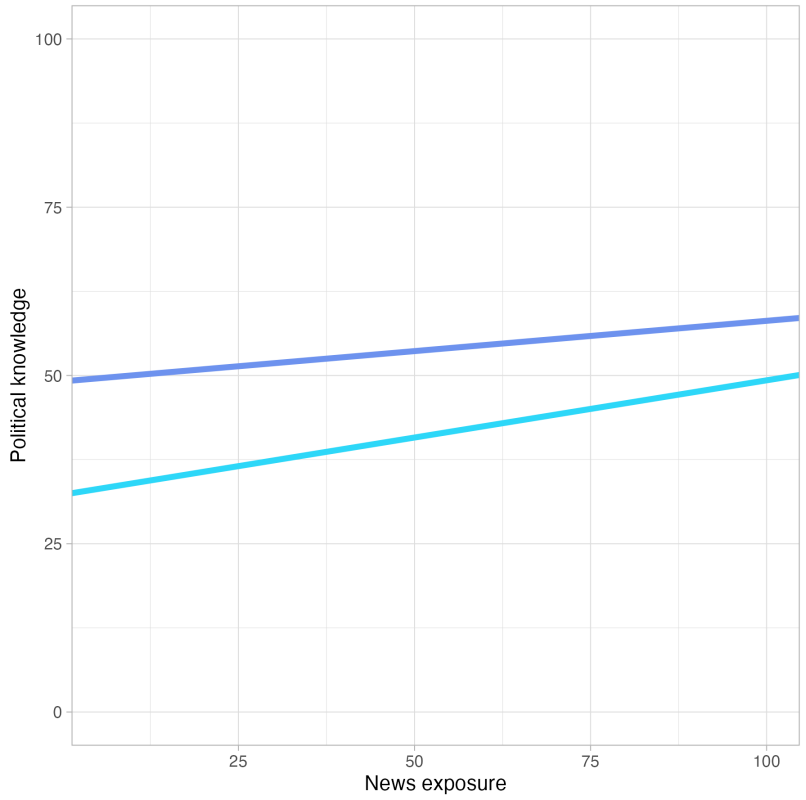
$$\widehat{\text{Knowledge}}_i = 32.28 + 0.17(\text{News Exposure}_i)$$

EDUCATION = 16 (UNDERGRADUATE EDUCATION)

$$\widehat{\text{Knowledge}}_i = -48.5 + 0.20(\text{Age}_i) + 4.21(\text{Education}_i) + 10.9(\text{Male}_i) + 0.24(\text{Engagement}_i) + -0.03(\text{Ideology}_i) + 6.51(\text{Democrat}_i) + 5.82(\text{Republican}_i) + 1.75(\text{Blue State}_i) + 0.41(\text{News Exposure}_i) - 0.02(\text{News Exposure}_i)(\text{Education}_i)$$

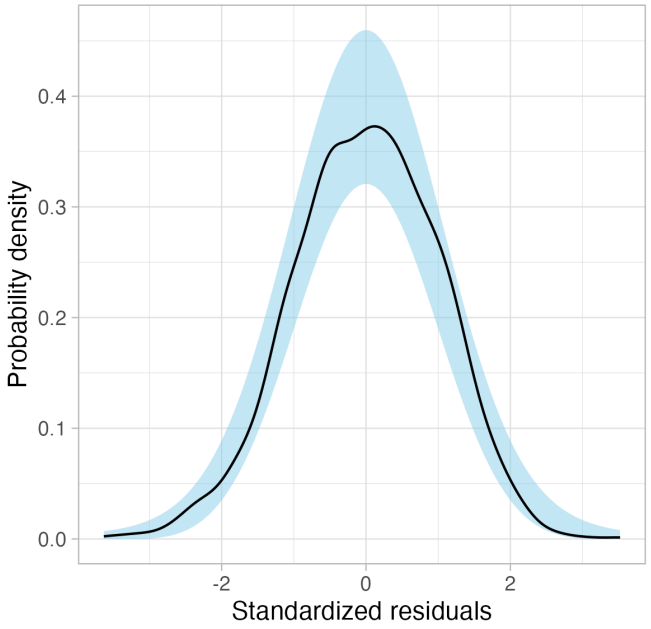
$$\widehat{\text{Knowledge}}_i = -48.5 + 0.20(50) + 4.21(16) + 10.9(0) + 0.24(50) - 0.03(0) + 6.51(1) + 5.82(0) + 1.75(1) + 0.41(\text{News Exposure}_i) - 0.02(\text{News Exposure}_i)(16)$$

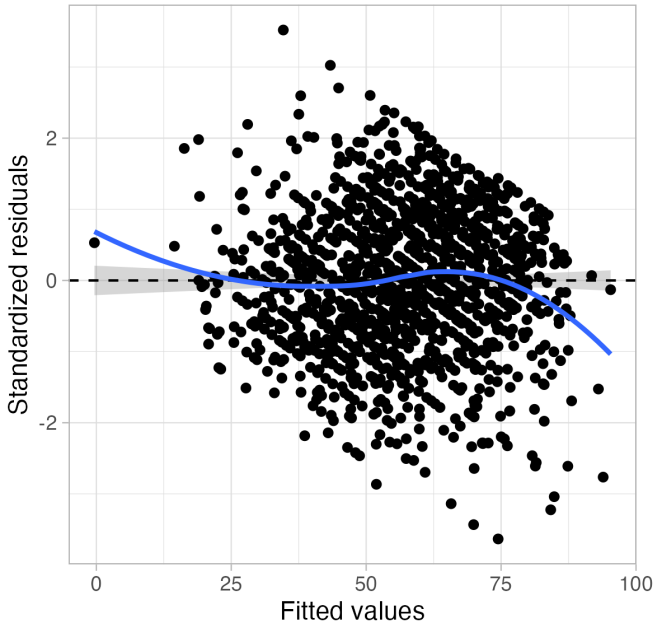
$$\widehat{\text{Knowledge}}_i = 49.12 + 0.09(\text{News Exposure}_i)$$



Predicted political knowledge as a function of news exposure for U.S. adult liberal female Democrats who are 50 years of age, are moderately engaged in politics, and live in a "blue" state. This is displayed for those with a **a high school** and **undergraduate** level of education.

RESIDUAL PLOTS FOR MODEL C





There still seems to be some violation of the linearity and homoscedasticity assumptions.

Table X. Unstandardized regression coefficients (and SEs) for a set of models predicting variation in political knowledge.

	Model A		Model B		Model C	
Predictor	B	SE	B	SE	B	SE
Age			0.20	0.03	0.20	0.03
Education level			3.42	0.20	4.21	0.43
Male			10.80	0.93	10.90	0.93
Political engagement			0.24	0.03	0.24	0.03
Political ideology			–0.03	0.02	–0.03	0.02
Democrat			6.53	1.43	6.51	1.43
Republican			5.85	1.38	5.82	1.38
Blue state			1.75	0.95	1.75	0.95
New Exposure	0.29	0.02	0.19	0.02	0.41	0.11
News exposure x Education level					–0.02	0.007
Intercept	42.4	1.25	–37.3	3.48	–48.5	6.35
RMSE	0.103		0.398		0.399	
R^2	21.7		17.8		17.8	

ANSWER RESEARCH QUESTIONS

There does seem to be a small, positive effect of news exposure on political knowledge.

After controlling for a set of demographic and political covariates, this effect persists, but diminishes slightly.

Education level, however, does seem to moderate this effect. The effect of news exposure on political knowledge is smaller for U.S. adults with higher levels of education, on average.

FOR NEXT CLASS: Consider your computer files and your organization of those files…

- Are your files organized into folders/directories? Or are they all in your Downloads folder?
- How did you organize all the data files, notes, etc. from EPsy 8251?
- If I asked you to find a specific file, could you locate it without using "Search"?
- Can you tell what is in a particular file by just looking at its name?
- Do your file names contain spaces? What about characters that aren't letters, numbers, dashes, or underscores?
- Are your file names consistent (all lower case letters, or all title case)? Or are they all different?