Statistical Reasoning

Week 9

Sciences Po - Louis de Charsonville

Spring 2018

Outline

Research Paper

Research Paper

Timeline

1^{st} **draft**	Done
No Class	3 April
2^{nd} **draft**	**10 April**
Week 11	17 April
Final draft	**24 April**

Single Regression

Simple regression by OLS

$$Y = \alpha + \beta X + \epsilon$$
$$\beta = \frac{Cov(X,Y)}{Var_X}$$
$$\alpha = \bar{Y} - \beta \bar{X}$$

- β is the estimate the variation in Y predicted by a change in one unit of X.
- The p-value test whether the coefficient is significantly different from 0.
- R^2 measures the *goodness of fit* and is the share of the variance of Y explain by the model.

Are CEO's wages correlated with sales ?

Are CEO's wages correlated with sales ?

- Model :

$$Wages = \alpha + \beta\, Sales + \epsilon$$

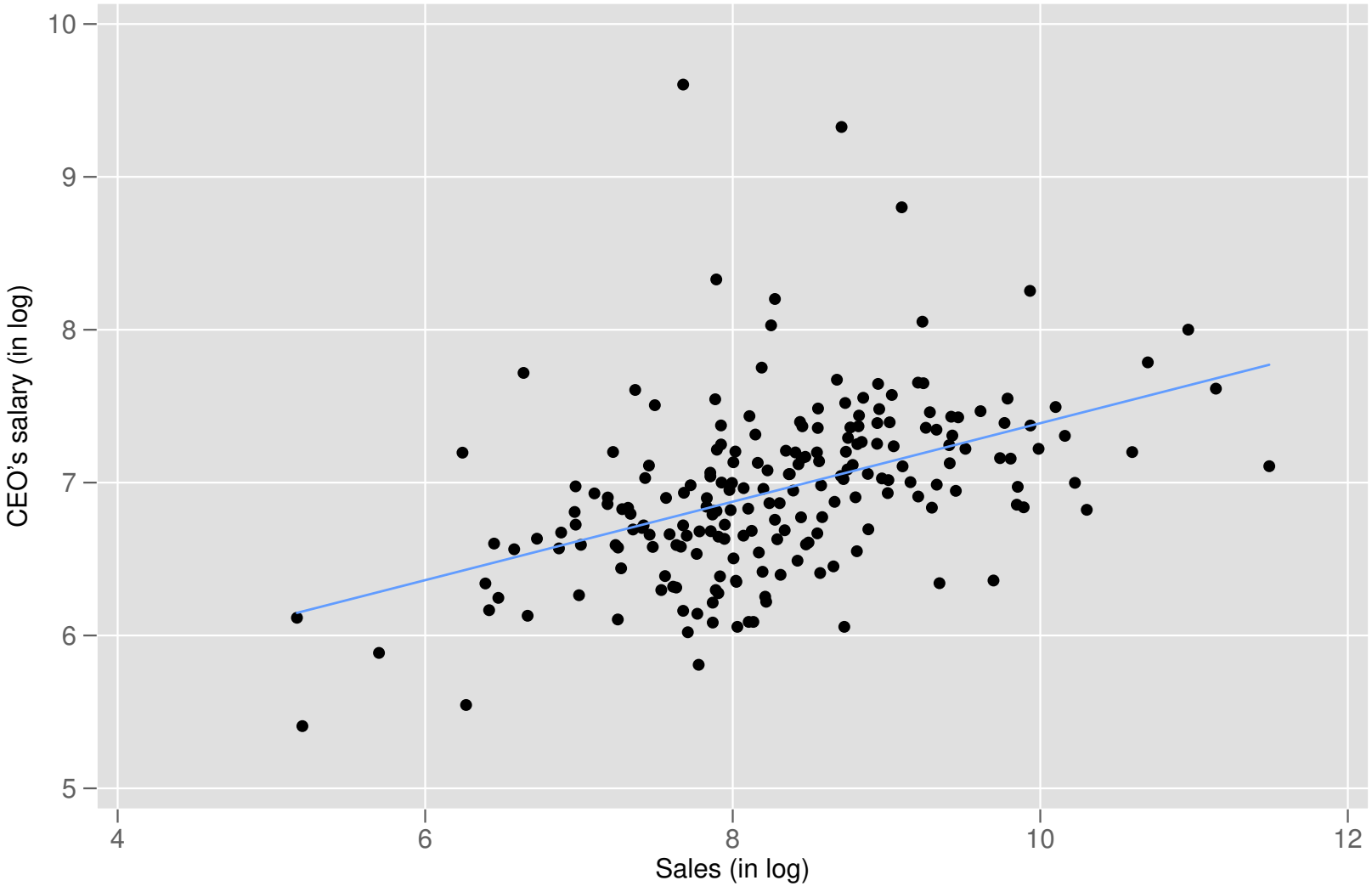
- in Stata :
  - Plot the data

```
tw (sc lsalary lsales) (lfit lsalary lsales)
```

  - Regression

```
reg lsalary lsales
```

Simple Regression - Reminder



```
. reg lsalary lsales
```

Source	SS	df	MS		
Model	14.0661688	1	14.0661688	Number of obs	= 209
				F(1, 207)	= 55.30
Residual	52.6559944	207	.254376785	Prob > F	= 0.0000
				R-squared	= 0.2108
				Adj R-squared	= 0.2070
Total	66.7221632	208	.320779631	Root MSE	= .50436

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsales	.2566717	.0345167	7.44	0.000	.1886224	.3247209
_cons	4.821997	.2883396	16.72	0.000	4.253538	5.390455

Hoes does the type of the firm impact the results ?

```
reg lsalary lsales if finance ==0
```

Source	SS	df	MS	Number of obs	=	163
				F(1, 161)	=	47.08
Model	12.4512191	1	12.4512191	Prob > F	=	0.0000
Residual	42.5750911	161	.26444156	R-squared	=	0.2263
				Adj R-squared	=	0.2215
Total	55.0263102	162	.339668582	Root MSE	=	.51424

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsales	.2584878	.0376703	6.86	0.000	.1840962	.3328795
_cons	4.782085	.3141753	15.22	0.000	4.161649	5.402521

```
reg lsalary lsales if finance ==1
```

Source	SS	df	MS	Number of obs	=	46
				F(1, 44)	=	6.49
Model	1.41543601	1	1.41543601	Prob > F	=	0.0144
Residual	9.60192044	44	.218225465	R-squared	=	0.1285
				Adj R-squared	=	0.1087
Total	11.0173565	45	.244830143	Root MSE	=	.46715

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lsales	.229666	.0901788	2.55	0.014	.0479226	.4114093
_cons	5.136137	.7576192	6.78	0.000	3.609256	6.663018

Multiple Regression

Multiple Linear Regression - Introduction

- First step into **Multivariate statistics**
- 1 dependent variable Y (should be continuous), multiple regressors $X_1, X_2, \ldots X_k$ (can be quantitative, ordinal)
- Can *control* the effect of X_i : disentangling effects of multiple independent variables.
- Determine which variable is the strongest predictor

Multiple Linear Regression - Model

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Partial derivatives

- Each coefficient is calculated by **holding all others constant** (*ceteris paribus*)
- It represents *net effects* (that's why control variables are so important).

Least squares

The model is still optimized by minimizing the squared error terms

Warning

The model is still assuming *linear*, *additive* relationships.

Does skipping lectures affect your educational attainment ?

- **Dependent Variable** : GPA score after graduation
- **Independent Variable** : Average nb of skipped lectures per week
- **Controls** :
  - High School GPA
  - Parents are college graduate
  - Has a personal computer
  - Gender
  - Age
  - Weekly Icool consumption

Stata

```
reg colGPA skipped hsGPA PC male age alcohol, beta
```

Source	SS	df	MS		
Model	5.26849772	6	.878082954	Number of obs	= 141
Residual	14.1376017	134	.10550449	F(6, 134)	= 8.32
				Prob > F	= 0.0000
Total	19.4060994	140	.138614996	R-squared	= 0.2715
				Adj R-squared	= 0.2389
				Root MSE	= .32481

colGPA	Coef.	Std. Err.	t	P>|t|	Beta
skipped	-.0765573	.0276927	-2.76	0.007	-.2239042
hsGPA	.4910405	.0911268	5.39	0.000	.4219506
PC	.1345645	.0575223	2.34	0.021	.1774824
male	.018962	.0598633	0.32	0.752	.0255246
age	.0262554	.0226128	1.16	0.248	.0896358
alcohol	.0309068	.0222811	1.39	0.168	.1141188
_cons	.7980128	.6400507	1.25	0.215	.

Dummies in a regression

Single coefficient

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (0) + \epsilon$$
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (1) + \epsilon$$

The omitted category $X_3 = 0$ is called the **reference category** and is part of the **baseline model** $Y = \alpha$, for which all coefficients are null.

Example

$$Income = \alpha + \beta_1 age + \beta_2 education + 0.male + \epsilon$$
$$Income = \alpha + \beta_1 age + \beta_2 education + 1.female + \epsilon$$

Categorical variables

Categorical variables can be used as **dummies**, e.g. binary recodes of each category that are tested agains a **reference category** to provide coefficients for the net effect of each category.

Stata

```
reg colGPA skipped hsGPA i.grad, beta
```

Source	SS	df	MS	Number of obs	=	141
				F(5, 135)	=	7.90
Model	4.39380249	5	.878760499	Prob > F	=	0.0000
Residual	15.012297	135	.1112022	R-squared	=	0.2264
				Adj R-squared	=	0.1978
Total	19.4060994	140	.138614996	Root MSE	=	.33347

colGPA	Coef.	Std. Err.	t	P>|t|	Beta
skipped	-.0795378	.0261501	-3.04	0.003	-.2326212
hsGPA	.4429045	.0910527	4.86	0.000	.3805873
grad					
2	.1133509	.1988868	0.57	0.570	.0440907
3	-.0290666	.0615775	-0.47	0.638	-.0391736
4	-.0096835	.1035004	-0.09	0.926	-.0075515
_cons	1.648639	.3241498	5.09	0.000	.

Radio and the Rise of the Nazis in Prewar Germany

(QJE 2015)

Adena, Enikolopov, Petrova, Santarosa, and Ekaterina Zhuravskaya

Motivation

- Dictators often come to power through a democratic process rather than military coups
  - Examples : Mugabe, Lukashenko, Chavez, Hitler
- How do future dictators persuade voters to support them ?
- When is propaganda more and less effective ?

Slides from material of E.Zhuravskaya

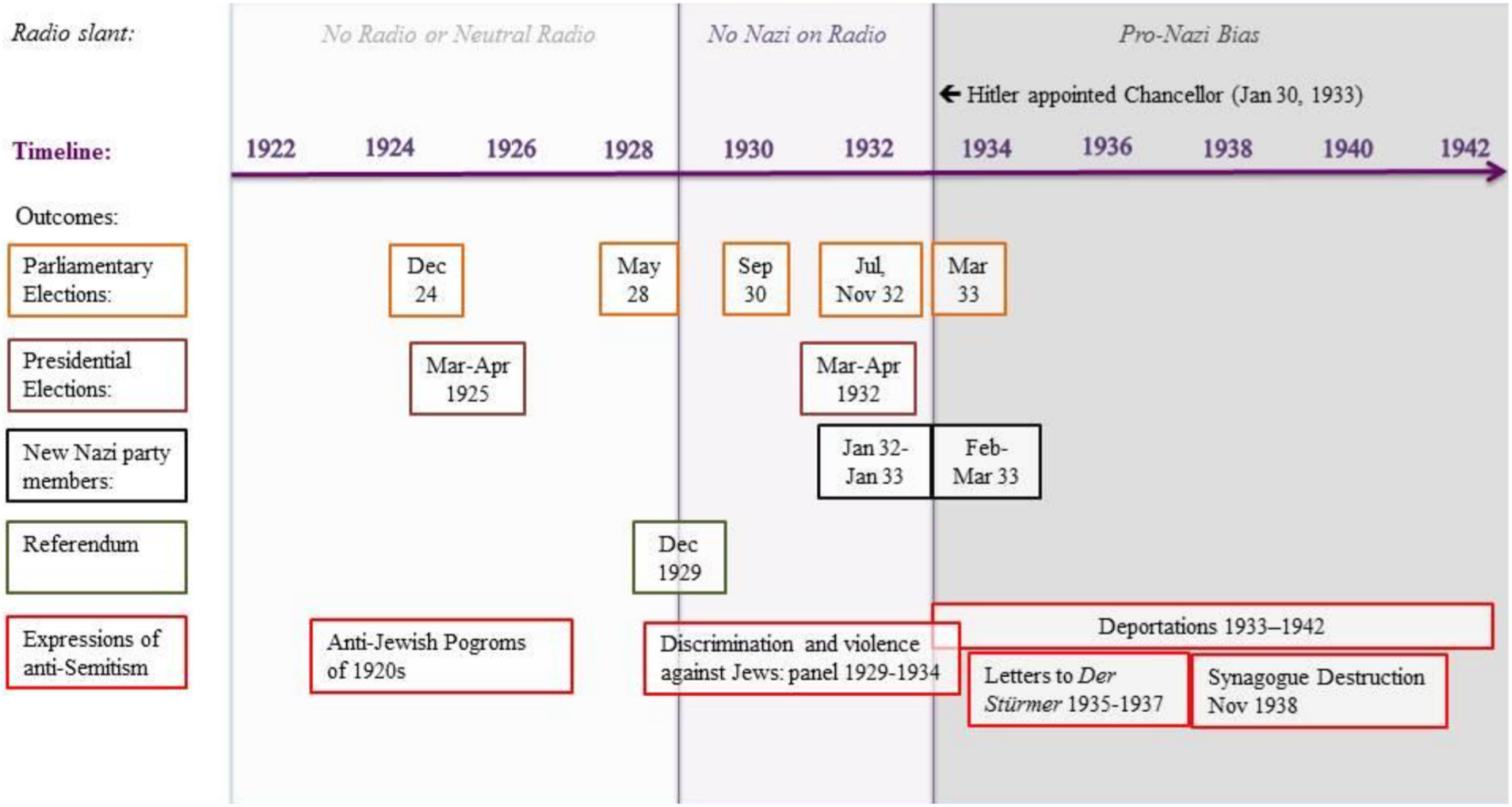
Main messages

- Whether future dictators or pro-democratic forces have control over mass media and whether extremist speech is allowed plays a role in preservation or collapse of immature democracies
- Propaganda can be very effective in maintaining popular support for dictator's policies, but it can also backfire and lead to lower support for the dictator
  - depending on listeners predisposition to the message

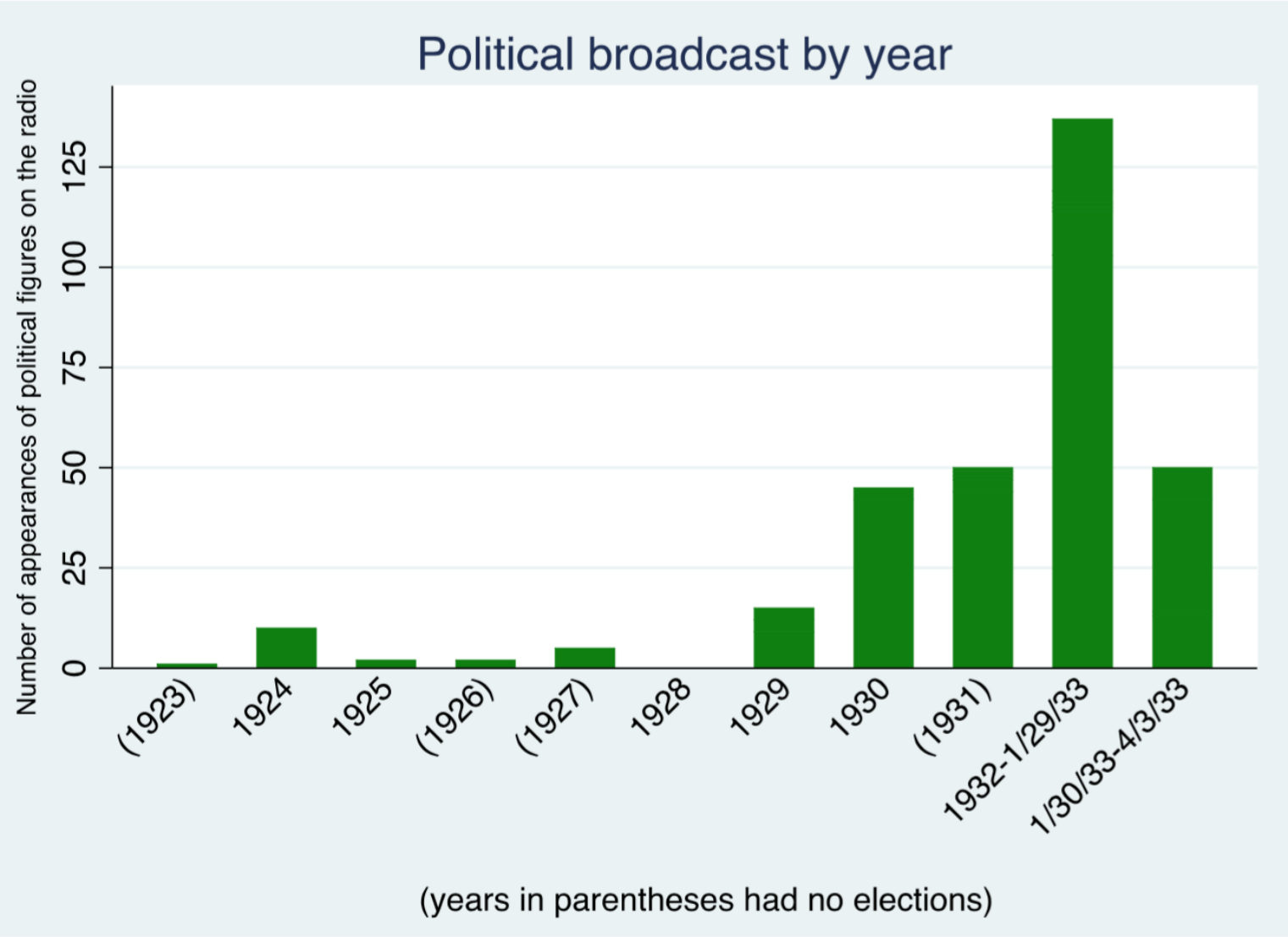
Why Nazi Germany ?

- The rise of the Third Reich is the most prominent example of a collapse of democracy without a military coup.
  - The Nazis won the March 1933 election (Nazi party got 43.9% of popular vote +8% for DNVP, their coalition partner) ; 18 days later parliament passed the Enabling Act.
- The Nazis themselves strongly believed in media power.
  - Aug 1933 : J.Goebbels "It would not have been possible for us to take power or to use it in the ways we have without the radio."

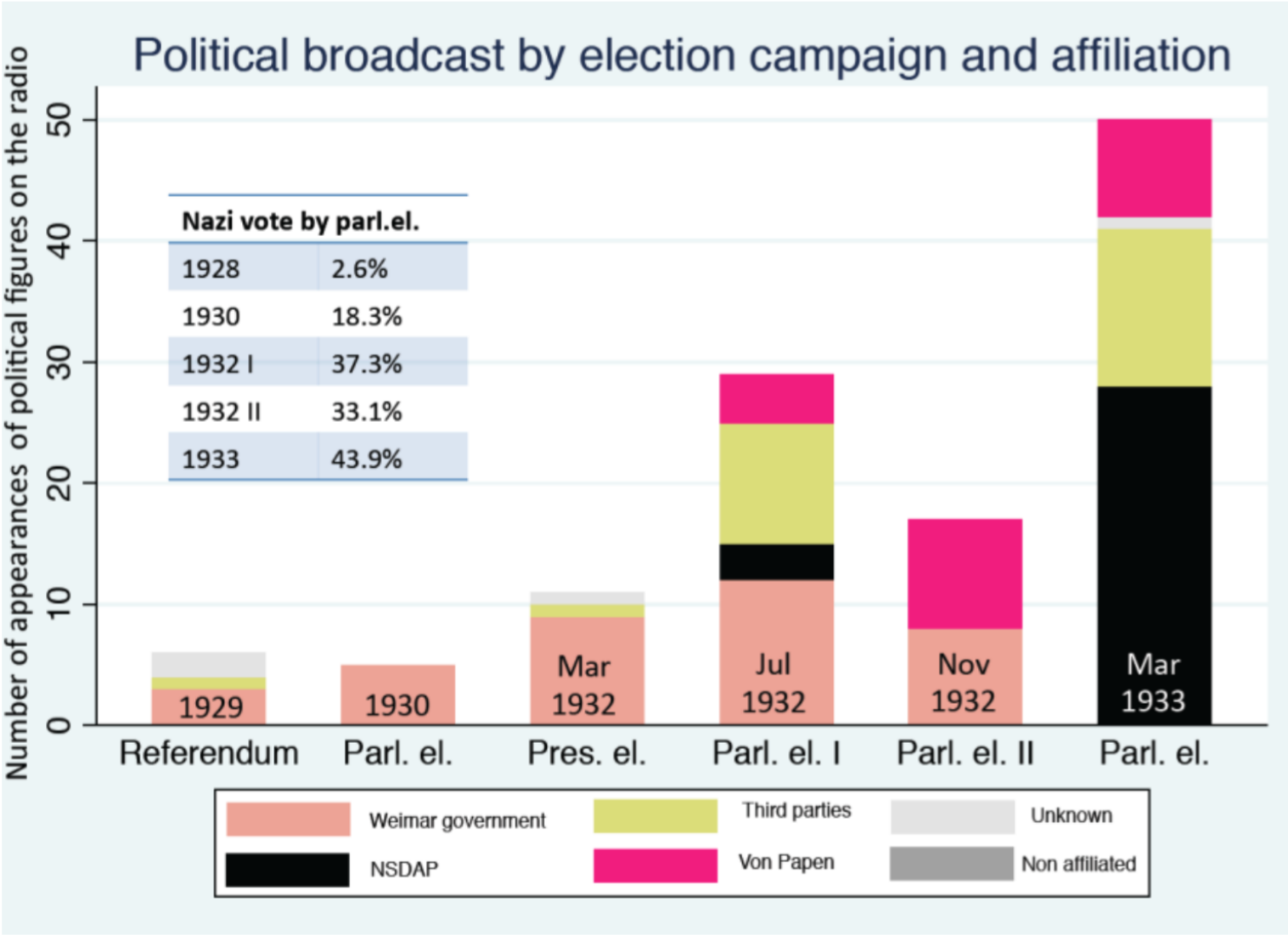
How does propaganda work ?



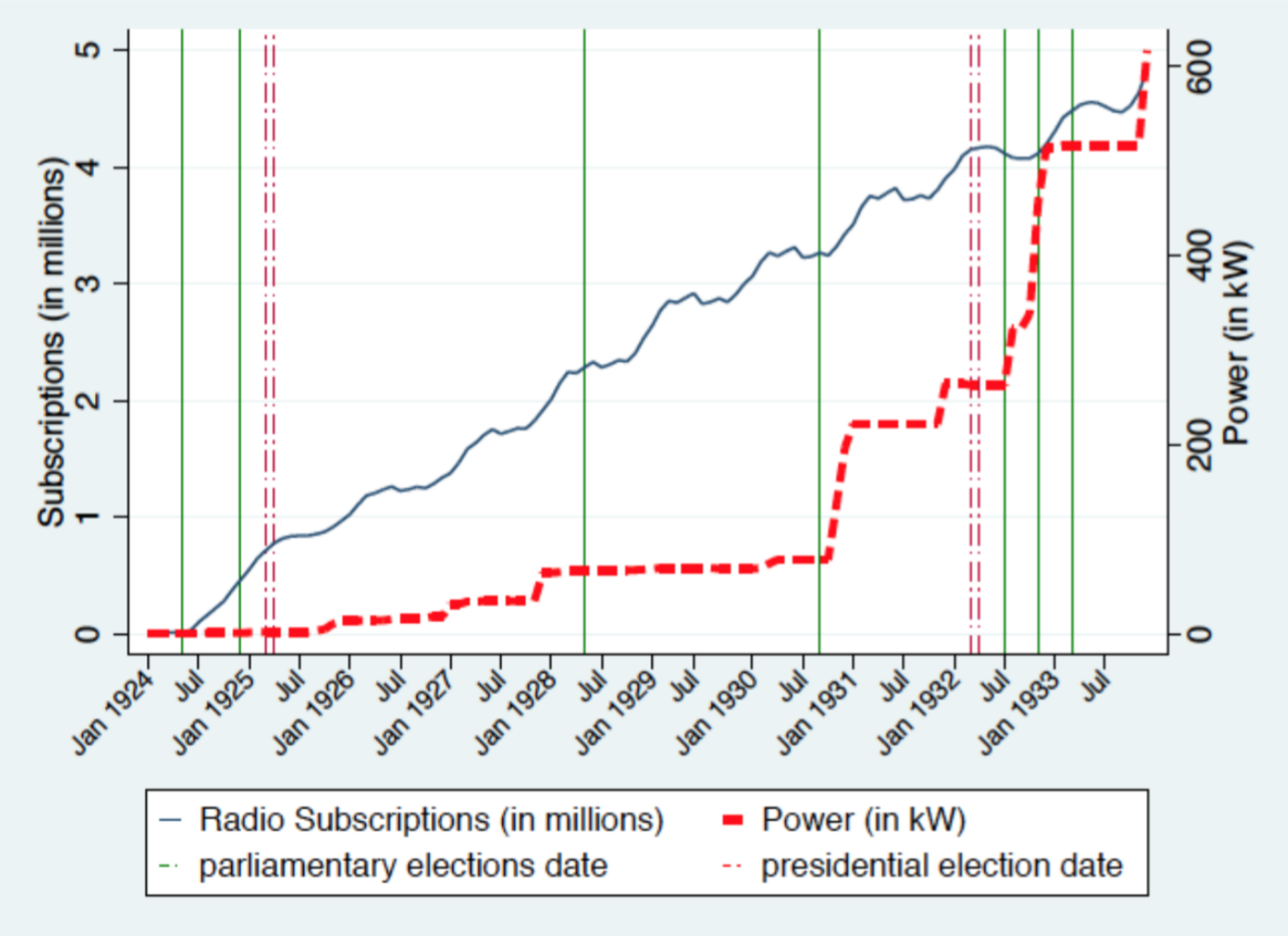
Radio became more and more political



Access to radio was unequal



Radio expanded quickly



Cross-section on first difference

$$\Delta y_{it} = \beta_{0t} + \beta_{1t} RadioExposure_{it} + \beta_{2t} X_{it} + \phi_p + \epsilon_t$$

With :

- y_{it} share of votes for the Nazis
- $RadioExposure_{it}$ signal strength
- X_{it} a vector of controls
  - Determinants of transmitter location
  - Socio-economic controls : census variables, including shares of Jews and Catholics, blue- and white-collar workers, WWI participation, property tax, welfare recipients
  - Voting preferences in 1924
  - Robust to controlling for newspapers, cinemas, and location of Hitler's speeches
- ϕ_p provinces fixed effects

Output

Panel A. Reduced form estimation

	Change in Vote Share of the Nazi Party Since Previous Elections			
Election dates:	*Sep 1930* (*Change from May 1928*)		*Mar 1933* (*Change from Nov 1932*)	
	(1)	(2)	(3)	(4)
Radio signal strength	-0.061***		0.045**	
	[0.022]		[0.020]	
Radio Signal Strength, non-linear transformation		-0.217***		0.128*
		[0.071]		[0.071]
Region fixed effects	Yes	Yes	Yes	Yes
Baseline controls	Yes	Yes	Yes	Yes
Observations	958	958	918	918

Research advices

Describe the overall model

- Total number of observations
- R-Squared
- *p*-value of the overall model (*F*-statistic)

Describe the coefficient fof your IVs

- *ceteris paribus*, what is the effect of x on y
- sign of the coefficient
- *p*-value of the coefficient
- Interpret the standardized coefficient, β, in order to compare the magnitude of each independent variable.

⚠ **Only the magnitude of the betas can be compared between independent variables, not the coefficients**

Summarizing a Multiple Linear Regression Model

"We ran a multiple regression analysis to examine the determinants of perception of the environment in France. Four predictors were included in the model : education, social class, trust in government, and age. Together, these factors account for 12% of the variance in environmental perceptions (R- Squared=0.12). All the variables except social class are signficiant (the p-values associated to the coefficients are lower than 0.05). Education and trust in government are the strongest predictors (beta=0.20) and are positively associated with environmental perceptions. Age is negatively related to environmental perceptions."

Descriptive and inferential bivariate statistics

Describe the Relationship

- Both variables are continuous `sc, pwcorr`
- If IV is a dummy : compare means `bysort, ttest`
- Both variables are categorical Cross-tabulations, Cramer's V

Signifiance

- Look at p-values of each kind of test (ttest, χ_2)

Don't confuse **Strengh** and **Significance**

- Cramer's V and Pearson's R are **not** statistical tests, but tell you the strength of the association ;
- Chi-2 and t-tests are **statistical tests** : they tell you whether the relationship is significant or not ;
- The `pwcorr` command with option `sig` or `star` provides both : Pearson's R and significance of the correlation ;
- In a regression model, the R-squared tells you the explanatory power of the predictor variables ;
- The *p*-values associated to the coefficients in the regression model tell you the statistical signifiance of the predictor.