

**POLI 30 D: Political Inquiry**  
Professor Umberto Mignozzetti  
(Based on DSS Materials)

Lecture 09 | Prediction II

## Before we start

### Announcements:

- ▶ Quizzes and Participation: On Canvas.
- ▶ GitHub page:  
<https://github.com/umbertomig/POLI30Dpublic>
- ▶ Piazza forum: Not sure what the link is. Ask your TA!
- ▶ Note to self: Turn on the mic!

## Before we start

**Recap:** We learned:

- ▶ The definitions of theory, scientific theory, and hypotheses.
- ▶ Data, datasets, variables, and how to compute means.
- ▶ Causal effect, treatments, outcomes, and randomization.
- ▶ Sampling, descriptive statistics, and descriptive plots for one variable.
- ▶ Correlation between two continuous variables.
- ▶ Prediction of a non-binary variable.

**Great job!**

- ▶ Do you have any questions about these contents?

## Plan for Today

- Prediction and Linear Regression
  - Example with Binary Outcome Variable:  
Using status quo Scores to Predict Probability of Supporting a Dictator
1. Load and explore data
  2. Identify X and Y
  3. What is the relationship between X and Y?
    - Create scatter plot
    - Calculate correlation
  4. Fit a linear model using the least squares method
  5. Interpret coefficients
  6. Make predictions
  7. Measure how well the model fits the data

## Predicting Support for a Dictator

- ▶ In 1988, FLACSO ran a survey to estimate the support for [Augusto Pinochet](#) in Chile.
- ▶ This survey was conducted on the eve of a referendum that could have ousted Pinochet.

---

variable	meaning
statusquo	Scale with status-quo evaluation. Roughly from -5 to 5.
vote	Declared vote in the upcoming referendum.
voteYES	1 means vote for Pinochet, and 0 means vote against it.

---

- ▶ We will study whether a person satisfied with the status quo would tend to favor Pinochet in the plebiscite.

## Step 1: Load and explore data

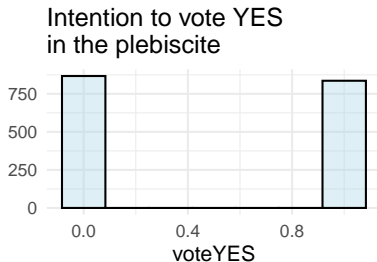
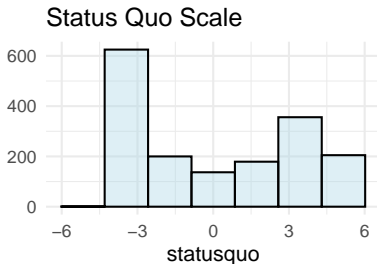
```
survchile <- read.csv("https://raw.githubusercontent.com/um  
head(survchile)
```

```
##      statusquo vote voteYES  
## 1      3.02460    Y        1  
## 2     -3.88851    N        0  
## 3      3.69216    Y        1  
## 4     -3.09489    N        0  
## 5     -3.31488    N        0  
## 6     -3.14055    N        0
```

- ▶ What is the unit of observation?
- ▶ For each variable: type and unit of measurement?
- ▶ Substantively interpret the first observation.

## Step 2: Identify the Dependent and Independent Variables

- ▶ The **predictor (X)** is the variable we want to use to predict the outcome (Y).
- ▶ The **target (Y)** is the variable that we want to predict.
- ▶ What are they?



## Step 2: Identify the Dependent and Independent Variables

- ▶ What type of variable is *voteYES*?
  - ▶ Binary
- ▶ How would you compute the proportion of intended Yes votes?
  - ▶ By computing the mean of *voteYES*
  - ▶ Since *voteYES* is a binary variable, its mean should be interpreted as the proportion of the observations that have the characteristic identified by the variable



## Step 2: Identify the Dependent and Independent Variables

- ▶ Code to compute the mean of *voteYES*
  - ▶ Answer:

```
mean(survchile$voteYES)  
## [1] 0.4908984
```

- ▶ Interpretation?
  - ▶ Close to 49.09% of people responded that they intended to support Pinochet in the upcoming plebiscite.
  - ▶ RECALL: You need to multiply the output by 100

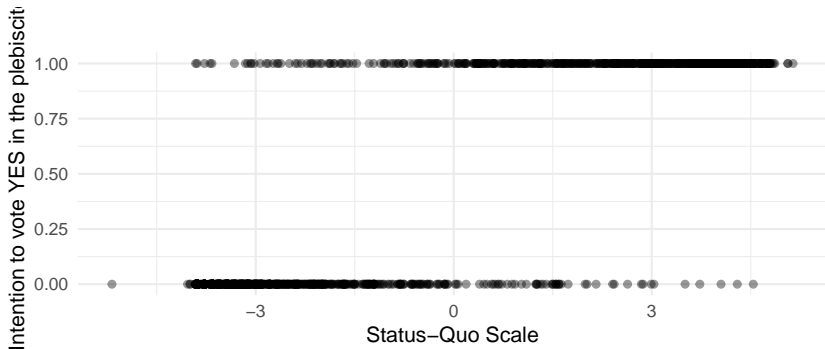
## Step 2: Identify the Dependent and Independent Variables

► Since Y is binary:

- unit of measurement of  $\bar{Y}$ ?
  - % (after x 100)
- unit of measurement of  $\hat{\beta}_0$ ?
  - % (after x 100)
- unit of measurement of  $\hat{Y}$ ?
  - % (after x 100)
- unit of measurement of  $\Delta \bar{Y}$ ?
  - p.p. (after x 100)
- unit of measurement of  $\hat{\beta}_1$ ?
  - p.p. (after x 100)
- unit of measurement of  $\Delta \hat{Y}$ ?
  - p.p. (after x 100)

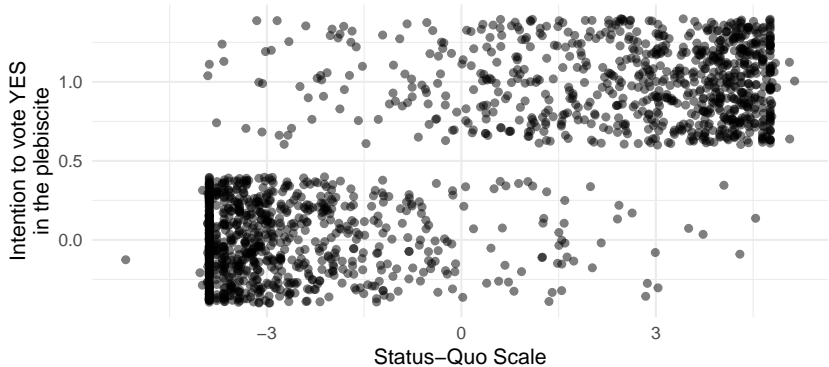
## Step 3: What is the relationship between X and Y?

- Create **scatter plot** to visualize the relationship between *statusquo* and *voteYES*.



- It is hard to see the y-axis variation. We add a little jitter on y, then.

### Step 3: What is the relationship between X and Y?



- What does each dot represent?
- Does the relationship look positive or negative?
- Does the relationship look weakly or strongly linear?

## Step 3: What is the relationship between X and Y?

- ▶ Calculate **correlation** to measure direction and strength of linear association between *statusquo* and *voteYES*

```
cor(survchile$statusquo, survchile$voteYES)  
## [1] 0.8535779
```

- ▶ We find a strong positive correlation
- ▶ Are we surprised by this?

## Step 4: Fit a linear model using the least squares method

- R function to fit a linear model: `lm()`

```
lm(voteYES ~ statusquo, data = survchile)
##
## Call:
## lm(formula = voteYES ~ statusquo, data = survchile)
##
## Coefficients:
## (Intercept)      statusquo
##      0.4927         0.1311
```

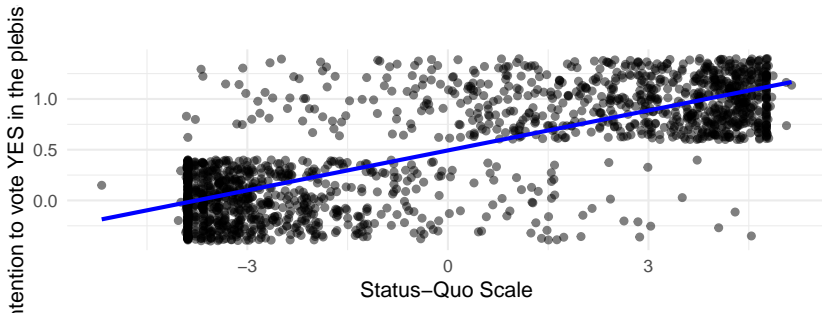
- $\hat{\beta}_0 = 0.49$  and  $\hat{\beta}_1 = 0.13$
- The fitted line is  $\hat{Y} = 0.49 + 0.13 X$
- More specifically, it is  $\widehat{\text{voteYES}} = 0.49 + 0.13 \text{ statusquo}$

## Step 4: Fit a linear model using the least squares method

- R function to add a fitted line to scatter plot:

`geom_smooth()`

```
ggplot(data = survchile, aes(x = statusquo, y = voteYES)) +  
  geom_jitter(fill = 'lightblue', alpha = 0.5, height = 0.4, width = 0) +  
  labs(title = '', y = 'Intention to vote YES in the plebiscite', x = 'Status-Quo Scale') +  
  geom_smooth(formula = 'y ~ x', method = 'lm', se = F, color = 'blue', lwd = 1) +  
  theme_minimal()
```



## Step 5: Interpretation of Coefficients

- ▶ Substantive interpretation of  $\hat{\beta}_0$ ?
  - ▶ Start with the mathematical definition:
    - ▶  $\hat{\beta}_0$  is the  $\hat{Y}$  when  $X=0$
  - ▶ Substitute  $X$ ,  $Y$ , and  $\hat{\beta}_0$ :
    - ▶  $\hat{\beta}_0 = 0.49$  is the  $\widehat{voteYES}$  when  $statusquo=0$
  - ▶ Put it in words (using units of measurement):
    - ▶ When a person is neither happy nor sad with things as they are, we predict that her probability of voting YES in the plebiscite is 49%, on average
- ▶ Unit of measurement of  $\hat{\beta}_0$ ?
  - ▶ Same as  $\bar{Y}$
  - ▶ Since  $Y$  is binary,  $\bar{Y}$  is measured in %, and so is  $\hat{\beta}_0$  (after  $\times 100$ )



## Step 5: Interpretation of Coefficients

- ▶ Substantive interpretation of  $\hat{\beta}_1$ ?
  - ▶ Start with the mathematical definition:
    - ▶  $\hat{\beta}_1$  is the  $\Delta \hat{Y}$  associated with  $\Delta X=1$
  - ▶ Substitute X, Y, and  $\hat{\beta}_1$ :
    - ▶  $\hat{\beta}_1 = 0.13$  is the  $\Delta \widehat{voteYES}$  associated with  $\Delta statusquo=1$
  - ▶ Put it in words (using units of measurement):
    - ▶ Increasing satisfaction with the status quo by 1 point is associated with a predicted increase in the chance of voting YES of 13 p.p., on average
- ▶ Unit of measurement of  $\hat{\beta}_1$ ?
  - ▶ Same as  $\Delta \bar{Y}$
  - ▶ Since Y is binary,  $\Delta \bar{Y}$  is measured in p.p., and so is  $\hat{\beta}_1$  (after x 100)

## Step 5: Interpretation of Coefficient

THE FITTED LINE IS

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- $\hat{\beta}_0$  (beta-zero-hat) is the estimated intercept coefficient  
the  $\hat{Y}$  when  $X=0$   
(in same unit of measurement as  $\bar{Y}$ )
- $\hat{\beta}_1$  (beta-one-hat) is the estimated slope coefficient  
the  $\Delta \hat{Y}$  associated with  $\Delta X=1$   
(in the same unit of measurement as  $\Delta \bar{Y}$ )

## Step 6: Make predictions

### USING THE FITTED LINE TO MAKE PREDICTIONS

- To predict  $\hat{Y}$  based on  $X$ :  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- To predict  $\Delta \hat{Y}$  based on  $\Delta X$ :  $\Delta \hat{Y} = \hat{\beta}_1 \Delta X$

## Step 6: Make predictions

To predict  $\hat{Y}$  based on  $X$ :  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

- Example 1: Imagine a person is unsatisfied with things and evaluates the status quo as -2. What would we predict her chance of favoring Pinochet in the plebiscite?

$$\widehat{\text{voteYES}} = 0.49 + 0.13 \text{ statusquo}$$

$$\widehat{\text{voteYES}} = 0.49 + 0.13 \times -2.0 \text{ (if statusquo} = -2.0\text{)}$$

$$\widehat{\text{voteYES}} = 0.23$$

- Answer: If her status quo evaluation is -2.0 points, we would predict that her probability of supporting Pinochet is of 23%, on average
- Note: since  $Y$  is binary,  $\hat{Y}$  is measured in % (after  $\times 100$ )

## Step 6: Make predictions

- Example 2: Imagine a person is happy with things and evaluates the status quo as 2. What would we predict her chance of favoring Pinochet in the plebiscite?

$$\widehat{\text{voteYES}} = 0.49 + 0.13 \text{ statusquo}$$

$$\widehat{\text{voteYES}} = 0.49 + 0.13 \times 2.0 \text{ (if statusquo} = 2.0\text{)}$$

$$\widehat{\text{voteYES}} = 0.75$$

- Answer: If the person scores 2.0 points on the status quo scale, we would predict that she would vote for Pinochet 75% of the time, on average

## Step 6: Make predictions

To predict  $\Delta \hat{Y}$  associated with  $\Delta X$ :  $\Delta \hat{Y} = \hat{\beta}_1 \Delta X$

- Example 3: If we raise a person's status quo evaluation by three points, how much would we predict that her support for Pinochet would change?

$$\Delta \widehat{\text{voteYES}} = 0.13 \Delta \text{statusquo}$$

$$\Delta \widehat{\text{voteYES}} = 0.13 \times 3.0 \quad (\text{if } \Delta \text{statusquo} = 3.0)$$

$$\Delta \widehat{\text{voteYES}} = 0.39$$

- Answer: An increase of status quo scores of 3 points is associated with a predicted increase in the probability of voting yes in the plebiscite of 39 p.p., on average.
- Note: Since Y is binary,  $\Delta \hat{Y}$  is in p.p. (after  $\times 100$ )

## Step 7: Measure how well the model fits the data with

- ▶ How good is the model at making predictions? How well does the model fit the data?
- ▶ One way of answering is by calculating  $R^2$

$R^2$  measures the proportion of the variation in the outcome variable explained by the model

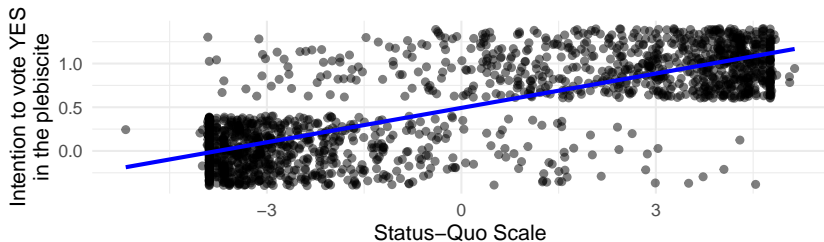
- ▶ It ranges from 0 to 1
- ▶ The higher the  $R^2$ , the better the model fits the data
- ▶ In the simple linear model:  $R^2 = \text{cor}(X, Y)^2$
- ▶ The higher the correlation between X and Y (in absolute terms), the better the model fits the data

## Step 7: Measure how well the model fits the data with $R^2$

- ▶ When  $\text{cor}(X,Y) = 1$  or  $\text{cor}(X,Y) = -1$ , the relationship between  $X$  and  $Y$  is perfectly linear.
- ▶  $R^2 = \text{cor}(X,Y)^2 = 1$ , the model explains 100% of the variation of  $Y$ .
- ▶ All prediction errors (vertical distance between the dots and the line) = 0.
- ▶ When  $\text{cor}(X,Y) = 0$ , the relationship between  $X$  and  $Y$  is non-linear.
- ▶  $R^2 = \text{cor}(X,Y)^2 = 0$ , the model explains 0% of the variation of  $Y$ .
- ▶ The prediction errors (vertical distance between the dots and the line) are large.



## Step 7: Measure how well the model fits the data with $R^2$



► Let us compute  $R^2$

```
cor(survchile$statusquo, survchile$voteYES)^2  
## [1] 0.7285953
```

## Step 7: Measure how well the model fits the data with $R^2$

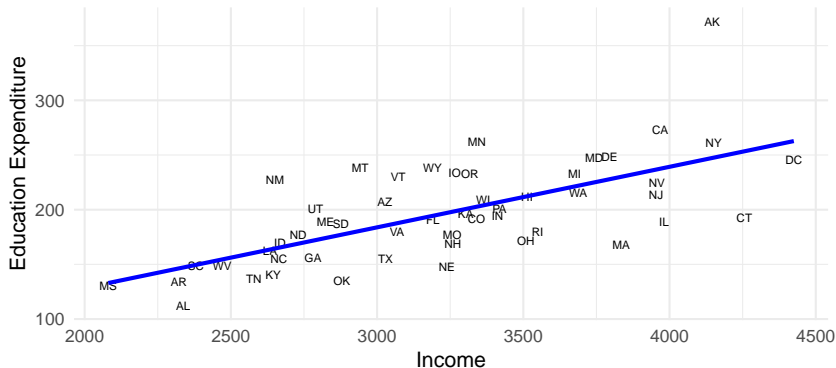
- Let us compute  $R^2$ :

```
cor(survchile$statusquo, survchile$voteYES)^2  
## [1] 0.7285953
```

- Interpretation?
  - It means that the linear model explains 73% of the variation of the outcome variable (*voteYES*)
  - **Note:** It does NOT mean that the model is right 73% of the time.

## Step 7: Measure how well the model fits the data with $R^2$

Let us return to the predictive model from the last lecture:



## Step 7: Measure how well the model fits the data with $R^2$

```
cor(educexp$income, educexp$education)^2  
## [1] 0.4456595
```

- Interpretation?
  - It means that the linear model explains 45% of the variation of the outcome variable (*education*)
  - It does NOT mean that the model is right 45% of the time
- Warnings:
  1. Only compare  $R^2$  between models with the same outcome variable ( $Y$ )
  2. Some variables are intrinsically harder to predict than others

## PREDICTING OUTCOMES USING LINEAR MODELS:

We look for  $X$  variables that are highly correlated with  $Y$  because the higher the correlation between  $X$  and  $Y$  (in absolute terms), the higher the  $R^2$  and the better the fitted linear model will usually be at predicting  $Y$  using  $X$ .

## Summary

### ► Today's Class:

- Practice summarizing the relationship between  $X$  and  $Y$  with a line: `lm()`.
- Practice interpreting the two estimated coefficients ( $\hat{\beta}_0$  and  $\hat{\beta}_1$ ) when outcome variable is binary.
- Practice making predictions with the fitted line: predict  $\hat{Y}$  based on  $X$  and predict  $\Delta\hat{Y}$  based on  $\Delta X$ .
- Learned how to measure how well the model fits the data with  $R^2$ .

### ► Next class:

- Causality with Observational Data

Questions?

See you in the next class!