POLI 30 D: Political Inquiry

Professor Umberto Mignozzetti
(Based on DSS Materials)

Lecture 08 | Prediction I

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!
- My mailbox disaster is not over, but things are in much better shape now! Please let me know if I missed your email.
- 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.

Great job!

- Do you have any questions about these contents?

Why Do We Analyze Data?

1. MEASURE: **To infer population characteristics via survey research**
   - what proportion of constituents support a particular policy?

2. PREDICT: **To make predictions**
   - who is the most likely candidate to win an upcoming election?

3. EXPLAIN: **To estimate the causal effect of a treatment on an outcome**
   - what is the effect of small classrooms on student performance?

Plan for Today

- Prediction and Linear Regression
- Example with Non-binary Target Variable:
  Use income to predict education expenditure

  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

1. **When estimating causal effects**

    - X is the **treatment** variable (independent variable)
    - Y is the **outcome** variable (dependent variable)
    - Aim: to estimate the effect of X on Y
    - Assumption: Treatment and control groups are comparable
    - Best way of satisfying assumption: random treatment assignment

2. **When infering population characteristics**

    - Aim: To infer the characteristics of X in the population
    - Assumption: sample is representative of population
    - Best way of satisfying assumption: Random sampling

3. **When making predictions**

- When we need to use what we know to learn what we do not know.
- X is a variable(s) that we use as predictor(s) (independent variable[s]; also k.a. **features**)
- Y is our **target variable**: what we want to predict
- $Y_i = f(X_i) + \varepsilon$; Where i is a given observation of interest; $f(.)$ is the *shape* of the relationship; and ε is the (inherent) error that the process entails.
- Aim: to predict Y as accurately as possible
- Assumption: The shape of f. We will assume linear: $f(X) = \beta_0 + \beta_1 X_i$.
- Best way to achieve our aim: To make R^2 as high as possible.

Using Income to Predict Education Expenditures is US States

- Today we will analyze data on 1970 U.S. State Public-School Expenditures.
- Our goal is to model the relationship between per-capita income and per-capita education expenditures.

Variable	Meaning
education	Per-capita education expenditures, dollars.
income	Per-capita income, dollars.
young	Proportion under 18, per 1000.
urban	Proportion urban, per 1000.
states	US State

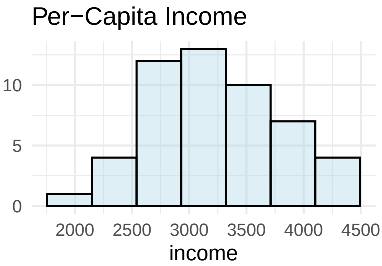
Step 1: Load and Explore Data

```r
grades <- read.csv("https://raw.githubusercontent.com/umber
educexp <- read.csv("https://raw.githubusercontent.com/umbe
head(educexp, 3)
##   education income young urban states
## 1       189   2824 350.7   508     ME
## 2       169   3259 345.9   564     NH
## 3       230   3072 348.5   322     VT
```

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

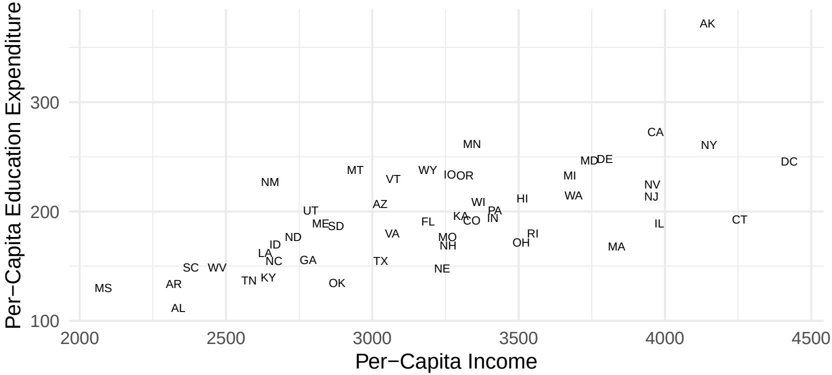
Step 2: Identify the Dependent and Independent Variables

- 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?

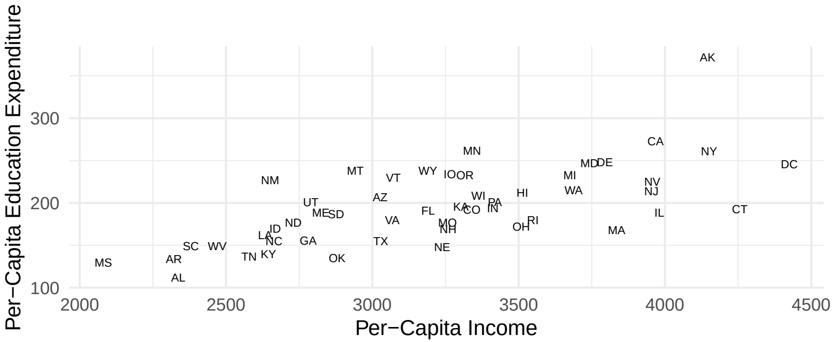


Step 3: What is the relationship between X and Y?

- Create scatter plot to visualize the relationship between per-capita *income* and *education* expenditures.



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



- The *Y variable* always goes in the *y-axis* and the *X variable* always goes in the *x-axis*.
- Does the relationship look positive or negative?
- Does the relationship look weekly or strongly linear?

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

- Let us now check the **correlation** coefficient.
- It measures the direction and strength of linear association between *income* and *education*.

```
cor(educexp$income, educexp$education)
## [1] 0.6675773
```

- We find a moderately strong positive correlation
- Are we surprised by this number? Think about what we have seen in the scatter plot.

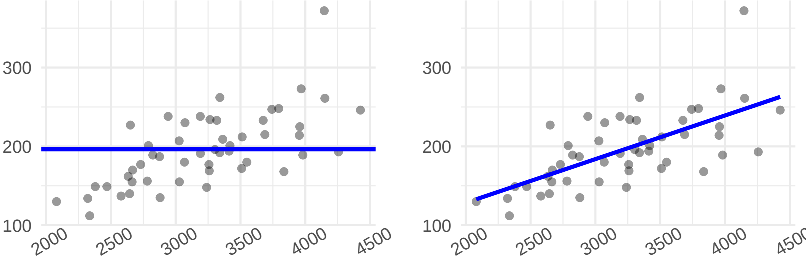
Step 3: What is the relationship between X and Y?

We learned so far:

- That an increase in per-capita *income* is associated with an increase in *education* expenditure.
- What we want to know is: When *income* increases, then **by how much** the *education* expenditure is predicted to increase?
- In general we care about: When X increases by one unit, by how much is Y predicted to change?
- To answer this question, we will fit a `regression` line to summarize the relationship between X and Y

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

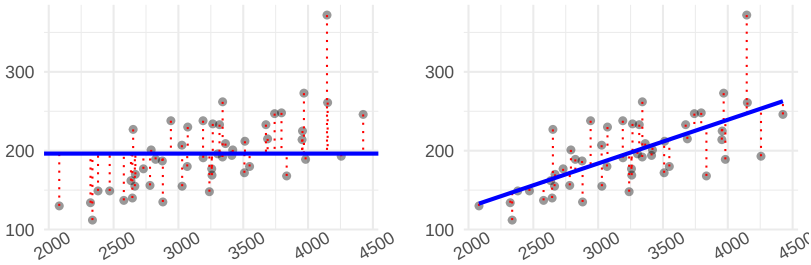
- Which line better summarizes the relationship?



- The goal is the choose the line that best fits the data.
  - Which one you think does that?

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

- To choose the line that best fits the data, we use **the least squares method**.
- In red, you can see the *error* we make by approximating the *education* using the blue trendline.



- Which plot you think is doing better?

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

- We need to think about what *better* means:
  - In the case of least square error, let the error in the prediction for a given US State i be:

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

- We need to find β_0 and β_1 that minimizes the sum of the squared error:

$$\min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} e_i^2 \quad \text{which is the same as} \quad \min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- The meaning of *least square method* should now be clear to you.

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

- The fitted line is $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1$ X
  - $\widehat{\beta}_0$ is the intercept
  - $\widehat{\beta}_1$ is the slope
- If you learned that a line was $Y = mX + b$
  - think that m is now $\widehat{\beta}_1$
  - think that b is now $\widehat{\beta}_0$
- ˆ (called 'hat') stands for predicted or estimated
  - $\widehat{Y}$ is the predicted target outcome
  - $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the estimated coefficients

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

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

```
lm('education ~ income', data = educexp)
##
## Call:
## lm(formula = "education ~ income", data = educexp)
##
## Coefficients:
## (Intercept)       income
##    17.71003      0.05538
```

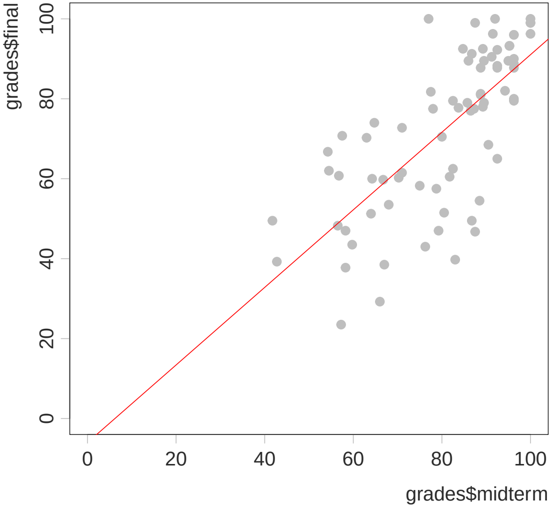
- $\widehat{\beta}_0 = 17.71$ and $\widehat{\beta}_1 = 0.06$
- The fitted line is $\widehat{Y} = 17.71 + 0.06\ X$
- More specifically: $\widehat{\text{education}} = 17.71 + 0.06\ \text{income}$

- We can now add the fitted line to the scatter plot above
- First, we store the fitted line in an object called *fit*

```
fit <- lm(grades$final ~ grades$midterm) # stores fitte
```

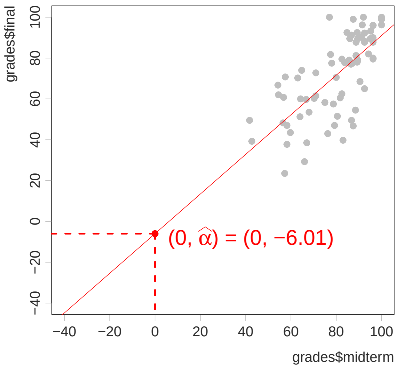
- Then, we can use the function **abline()**
  - required argument: name of object with fitted line

```
abline(fit) # adds line to scatter plot
```


grades$final
0
20
40
60
80
100
0
20
40
60
80
100
grades$midterm

5. Interpretation of Coefficients:

- **The intercept ($\widehat{\alpha}$) is the $\widehat{Y}$ when X=0**
- Find 0 on the X-axis, go up to the line, find the value of $\widehat{Y}$ associated with X=0



- here: $\widehat{\alpha}$ = -6.01

Mathematical definition of $\widehat{\alpha}$

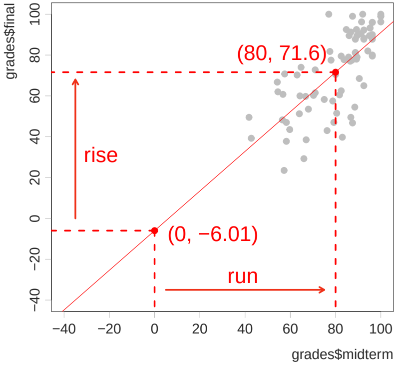
$$
\begin{aligned}
\widehat{Y} &= \widehat{\alpha} + \widehat{\beta}\, X && \text{(by definition)} \\
\widehat{Y} &= \widehat{\alpha} + \widehat{\beta} \times 0 && \text{(if } X = 0\text{)} \\
\widehat{Y} &= \widehat{\alpha} + 0 && \text{(if } X = 0\text{)} \\
\widehat{Y} &= \widehat{\alpha} && \text{(if } X = 0\text{)}
\end{aligned}
$$

$\widehat{\alpha}$ *is the value of* $\widehat{Y}$ *when* $X=0$

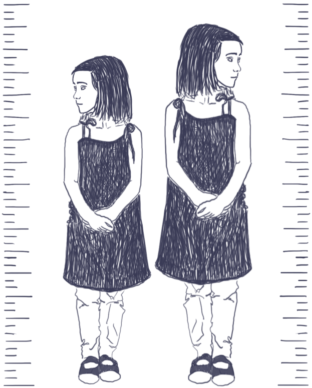
- **substantive interpretation of $\widehat{\alpha}$?**
  - start with mathematical definition:
    - $\widehat{\alpha}$ is the $\widehat{Y}$ when X=0
  - substitute X, Y, and $\widehat{\alpha}$:
    - $\widehat{\alpha}$ = -6.01 is the $\widehat{\textit{final}}$ when *midterm*=0
  - put it in words (using units of measurement):
    - when a student scores 0 points in the midterm, we predict that in the final exam they will score -6.01 points, on average
  - sometimes it is nonsensical (due to extrapolation)
- **unit of measurement of $\widehat{\alpha}$?**
  - same as $\overline{Y}$
  - in this case: Y is non-binary and measured in points so $\overline{Y}$ is measured in points and so is $\widehat{\alpha}$

5. Interpretation of Coefficients:

- **The slope ($\widehat{\beta}$) is the $\triangle \widehat{Y}$ associated with $\triangle X$=1**
- Pick two points on the line, measure $\triangle \widehat{Y}$ and $\triangle X$ associated with the two points, calculate $\triangle \widehat{Y} / \triangle X$



- here: $\widehat{\beta} = \frac{\text{rise}}{\text{run}} = \frac{71.6-(-6.01)}{80-0} = \frac{77.61}{80} = 0.97$



$$\triangle \widehat{Y} = \widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}}$$
$$\triangle X = X_{\text{final}} - X_{\text{initial}}$$

Mathematical definition of $\widehat{\beta}$

$$\begin{aligned}
\triangle \widehat{Y} &= \widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}} && \text{(by definition)} \\
\triangle \widehat{Y} &= (\widehat{\alpha} + \widehat{\beta} X_{\text{final}}) - (\widehat{\alpha} + \widehat{\beta} X_{\text{initial}}) && \text{(since } \widehat{Y}_{\text{final}} = \widehat{\alpha} + \widehat{\beta} X_{\text{final}} \\
& && \text{and } \widehat{Y}_{\text{initial}} = \widehat{\alpha} + \widehat{\beta} X_{\text{initial}}) \\
\triangle \widehat{Y} &= \widehat{\alpha} - \widehat{\alpha} + \widehat{\beta}(X_{\text{final}} - X_{\text{initial}}) && \text{(rearranging terms)} \\
\triangle \widehat{Y} &= \widehat{\beta}(X_{\text{final}} - X_{\text{initial}}) && \text{(since } \widehat{\alpha} - \widehat{\alpha} = 0) \\
\triangle \widehat{Y} &= \widehat{\beta}(\triangle X) && \text{(since } \triangle X = X_{\text{final}} - X_{\text{initial}}) \\
\triangle \widehat{Y} &= \widehat{\beta} \times 1 && \text{(if } \triangle X = 1) \\
\triangle \widehat{Y} &= \widehat{\beta} && \text{(if } \triangle X = 1)
\end{aligned}$$

$\widehat{\beta}$ is the value of $\triangle \widehat{Y}$ associated with $\triangle X = 1$

- **substantive interpretation of $\widehat{\beta}$?**
  - start with mathematical definition:
    - $\widehat{\beta}$ is the $\triangle\widehat{Y}$ associated with $\triangle X$=1
  - substitute X, Y, and $\widehat{\beta}$:
    - $\widehat{\beta} = 0.97$ is the $\triangle\widehat{\mathit{final}}$ associated with $\triangle \mathit{midterm}$=1
  - put it in words (using units of measurement):
    - an increase in midterm scores of 1 point is associated with a predicted increase in final exam scores of 0.97 points, on average
  - has the same sign as the cor(X,Y) (always the case!)
- **unit of measurement of $\widehat{\beta}$?**
  - same as $\triangle\overline{Y}$
  - in this case: Y is non-binary and measured in points so $\triangle\overline{Y}$ is measured in points and so is $\widehat{\beta}$

THE FITTED LINE IS:

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$$

- $\widehat{\alpha}$ (alpha-hat) is the estimated intercept coefficient
  the $\widehat{Y}$ when X=0
  (in same unit of measurement as $\overline{Y}$)

- $\widehat{\beta}$ (beta-hat) is the estimated slope coefficient
  the $\triangle\widehat{Y}$ associated with $\triangle X$=1
  (in the same unit of measurement as $\triangle\overline{Y}$)

6. Make predictions

- Now that we have found the line that best summarizes the relationship between X and Y, we can use it to make predictions

- There are two types of predictions that we might be interested in:

    1. predict $\widehat{Y}$ based on X: $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$
    2. predict $\triangle\widehat{Y}$ associated with $\triangle X$: $\triangle\widehat{Y} = \widehat{\beta}\triangle X$

To predict $\widehat{Y}$ based on X: $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$

- Example 1: Imagine you earn 80 points in the midterm, what would we predict your final exam score will be?

$$\widehat{\text{final}} = \text{-}6.01 + 0.97 \;\; \text{midterm}$$
$$\widehat{\text{final}} = \text{-}6.01 + 0.97 \times 80 \;\; (\text{if midterm} = 80)$$
$$\widehat{\text{final}} = 71.59$$

- Answer: If you earn 80 points in the midterm, we would predict that you will earn 71.59 points in the final exam, on average
- Note: $\widehat{Y}$ is in the same unit of measurement as $\overline{Y}$
  - in this case: Y is non-binary and measured in points so $\overline{Y}$ and $\widehat{Y}$ are also measured in points

- Example 2: Imagine you earn 90 points in the midterm, what would we predict your final exam score will be?

$$\begin{aligned}
\widehat{\text{final}} &= \text{-}6.01 + 0.97 \;\; \text{midterm} \\
\widehat{\text{final}} &= \text{-}6.01 + 0.97 \times 90 \;\; (\text{if midterm} = 90) \\
\widehat{\text{final}} &= 81.29
\end{aligned}$$

- Answer: If you earn 90 points in the midterm, we would predict that you will earn 81.29 points in the final exam, on average

To predict $\triangle\widehat{Y}$ associated with $\triangle X$: $\triangle\widehat{Y} = \widehat{\beta}\,\triangle X$

- Example 3: If you increase your midterm scores by 10 points, by how much would we predict that your final exam scores would change?

$$\begin{aligned}\triangle\widehat{\text{final}} &= 0.97\ \triangle\text{midterm}\\ \triangle\widehat{\text{final}} &= 0.97 \times 10 \quad (\text{if } \triangle\text{midterm} = 10)\\ \triangle\widehat{\text{final}} &= 9.7\end{aligned}$$

- Answer: An increase of midterm scores of 10 points is associated with a predicted increase of final exam scores of 9.7 points, on average
- Note: $\triangle\widehat{Y}$ is in the same unit of measurement as $\triangle\overline{Y}$
  - in this case: Y is non-binary and measured in points so $\triangle\overline{Y}$ and $\triangle\widehat{Y}$ are also measured in points

7. Measure how well the model fits the data with R^2

- We will see how to do this next lecture

Today's Class

- How to summarize the relationship between X and Y with a line: `lm()` and `abline()`
- How to interpret the two estimated coefficients: ($\widehat{\alpha}$ and $\widehat{\beta}$) when outcome variable is non-binary
- How to make predictions with the fitted line: predict $\widehat{Y}$ based on X and predict $\triangle\widehat{Y}$ based on $\triangle X$

Next Class

- Another example of how to use the linear model to make predictions, but with binary outcome
- How to measure how well the model fits the data with R^2

For Next Class

Here is the friendly reminder of what is due for next class:

- Go see the peer tutors (plan ahead)
- Review what you have learned thus far (lectures 2-10)
- Do new set of readings: section 4.6-4.9 (including both, following along the exercises with your own computer, skip 4.8)
- Bring your course packets to class

Summary

- **Today's Class:**
  - Exploring the Relationship Between Two Variables
    - Scatterplots
    - Correlations
- Next class:
  - Prediction and Linear Regression

Questions?

See you in the next class!