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 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 inferring population characteristics**
    - Aim: To infer the characteristics of X in the population
    - Assumption: sample is representative of the 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 in 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

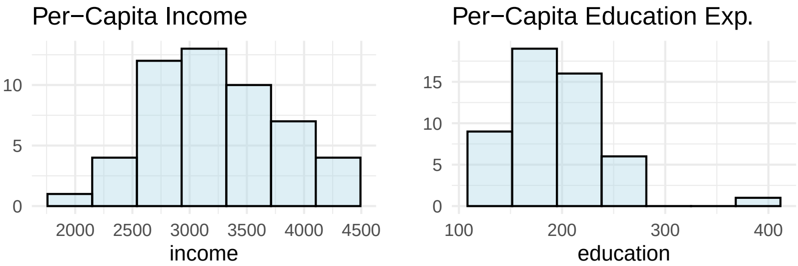
Step 1: Load and Explore Data

```r
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

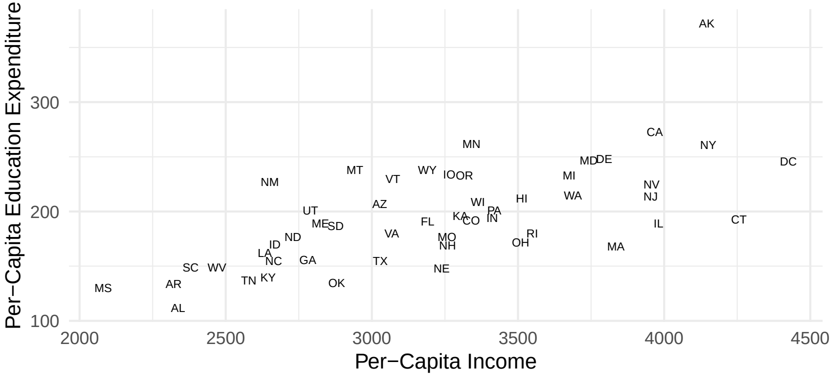
Step 2: Identify the Dependent and Independent Variables

- 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?

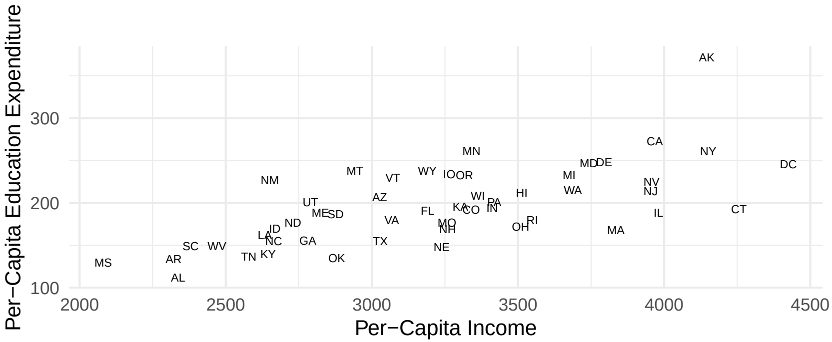


Step 3: What is the relationship between X and Y?

- 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 the 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.

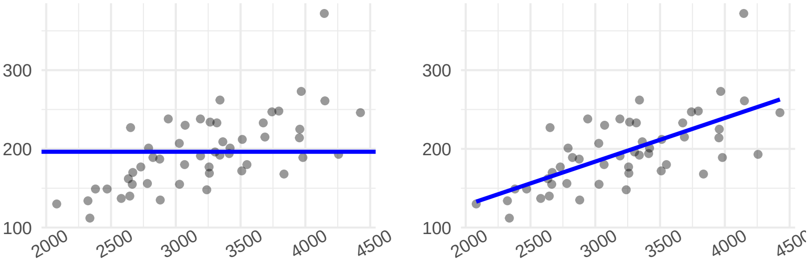
Step 3: What is the relationship between X and Y?

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

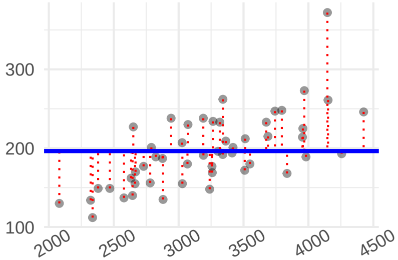
- Which line better summarizes the relationship?

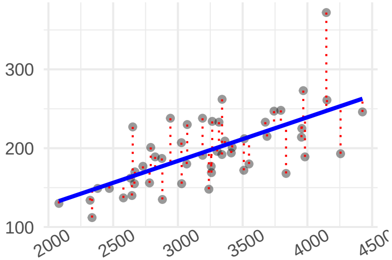


- The goal is the choose the line that best fits the data.
  - Which one do you think does that?

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

- To choose the line 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 do 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()`:

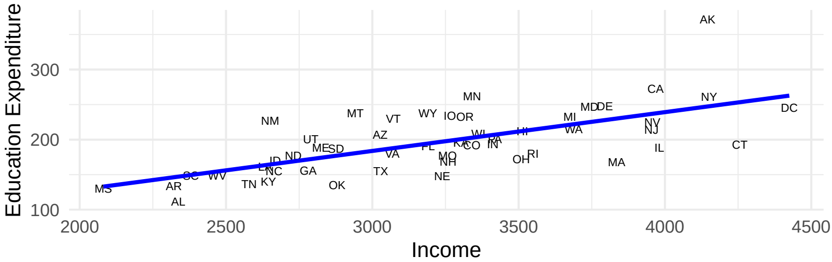
```
lm('education ~ income', data = educexp)
##
## Call:
## lm(formula = "education ~ income", data = educexp)
##
## Coefficients:
## (Intercept)       income
##    17.71003      0.05538
```

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

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

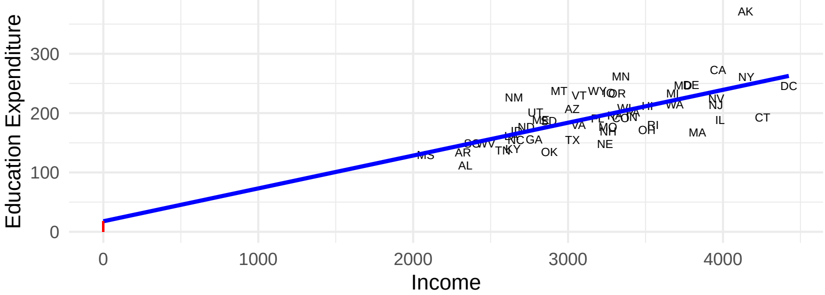
- And to add a fitting line to a scatter plot, you can use `abline()` or `geom_smooth()`.
  - You are going to learn this in Labs 06 and 07.

```
ggplot(data = educexp, aes(x = income, y = education)) + geom_text(aes(label=states), size=2) +
  labs(title = '', y = 'Education Expenditure', x = 'Income') +
  geom_smooth(formula = 'y ~ x', method = 'lm', se = F, color = 'blue', lwd = 1) + theme_minimal()
```



Step 5: Interpretation of Coefficients

- **The intercept ($\widehat{\beta_0}$) is the $\widehat{Y}$ when X=0.**



- here: $\widehat{\beta_0} = 17.71$

Mathematical definition of $\widehat{\beta}_0$

$$
\begin{aligned}
\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \, X && \text{(by definition)} \\
\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \times 0 && \text{(if } X = 0\text{)} \\
\widehat{Y} &= \widehat{\beta}_0 + 0 && \text{(if } X = 0\text{)} \\
\widehat{Y} &= \widehat{\beta}_0 && \text{(if } X = 0\text{)}
\end{aligned}
$$

$\widehat{\beta}_0$ is the value of $\widehat{Y}$ when $X = 0$

Substantive interpretation of $\widehat{\beta}_0$

- Substitute X, Y, and $\widehat{\beta}_0$:
  - $\widehat{\beta}_0 = 17.71$ is the $\widehat{education}$ when income $= 0$
  - When a State has 0 per-capita income, we predict that the per-capita expenditure in education will be 17.71 dollars, on average
    - Sometimes, it is nonsensical (due to extrapolation)

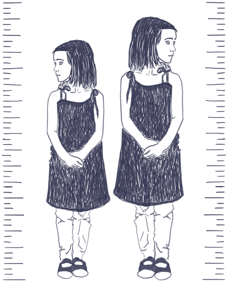
- **Unit of measurement of** $\widehat{\beta}_0$?
  - Same as $\overline{Y}$
  - In this case: Y is non-binary and measured in points so $\overline{Y}$ is measured in points and so is $\widehat{\beta}_0$

Step 5: Interpretation of Coefficients

- 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: $\widehat{\beta}_1 = 0.06$

Step 5: Interpretation of Coefficients

- The slope ($\widehat{\beta_1}$) is the $\triangle\widehat{Y}$ associated with $\triangle X$=1



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

Mathematical definition of $\widehat{\beta}_1$

$$\triangle\widehat{Y} = \widehat{Y}_{\text{final}} - \widehat{Y}_{\text{initial}} \qquad \text{(by definition)}$$

$$\triangle\widehat{Y} = (\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{final}}) - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{initial}}) \qquad (\text{since } \widehat{Y}_{\text{final}} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{final}} \text{ and } \widehat{Y}_{\text{initial}} = \widehat{\beta}_0 + \widehat{\beta}_1 X_{\text{initial}})$$

$$\triangle\widehat{Y} = \widehat{\beta}_0 - \widehat{\beta}_0 + \widehat{\beta}_1 (X_{\text{final}} - X_{\text{initial}}) \qquad \text{(rearranging terms)}$$

$$\triangle\widehat{Y} = \widehat{\beta}_1 (X_{\text{final}} - X_{\text{initial}}) \qquad (\text{since } \widehat{\beta}_0 - \widehat{\beta}_0 = 0)$$

$$\triangle\widehat{Y} = \widehat{\beta}_1 (\triangle X) \qquad (\text{since } \triangle X = X_{\text{final}} - X_{\text{initial}})$$

$$\triangle\widehat{Y} = \widehat{\beta}_1 \times 1 \qquad (\text{if } \triangle X = 1)$$

$$\triangle\widehat{Y} = \widehat{\beta}_1 \qquad (\text{if } \triangle X = 1)$$

$\widehat{\beta}_1$ is the value of $\triangle\widehat{Y}$ associated with $\triangle X = 1$

Substantive interpretation of $\widehat{\beta}_1$

- Start with the mathematical definition:
  - $\widehat{\beta}_1$ is the $\triangle\widehat{Y}$ associated with $\triangle X$=1
- Substitute X, Y, and $\widehat{\beta}_1$:
  - $\widehat{\beta}_1$ = 0.06 is the $\triangle\widehat{education}$ associated with $\triangle income$=1
    - An increase in income of 1 dollar is associated with a predicted increase in education expenditures of 6 cents, on average
- **Unit of measurement of** $\widehat{\beta}_1$?
  - 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 $\widehat{\beta}_1$

THE FITTED LINE IS:

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

- $\widehat{\beta}_0$ (beta-zero-hat) is the estimated intercept coefficient
  the $\widehat{Y}$ when $X=0$
  (in same unit of measurement as $\overline{Y}$)

- $\widehat{\beta}_1$ (beta-one-hat) is the estimated slope coefficient
  the $\triangle\widehat{Y}$ associated with $\triangle X=1$
  (in the same unit of measurement as $\triangle\overline{Y}$)

Step 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: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$

2. predict $\triangle\widehat{Y}$ associated with $\triangle X$: $\triangle\widehat{Y} = \widehat{\beta}_1 \triangle X$

Step 6: Make Predictions

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

- Example 1: Suppose we are back in the 70s. Imagine you lived in a State where the per-capita income is $ 3,500. What would the education expenditure be?

$$\begin{aligned}
\widehat{\text{education}} &= 17.71 + 0.06 \ \ \text{income} \\
\widehat{\text{education}} &= 17.71 + 0.06 \times 3{,}500 \quad (\text{if income} = 3{,}500) \\
\widehat{\text{education}} &= 227.71
\end{aligned}$$

- Answer: If the income per-capita was $ 3,500, the the education expenditure would be $ 227.71, on average.

Step 6: Make Predictions

To predict $\triangle \widehat{Y}$ associated with $\triangle X$: $\triangle \widehat{Y} = \widehat{\beta}_1 \triangle X$

- Example 2: Suppose the per-capita income rises by \$100. By how much would we predict that the education expenditure would change?

$$
\begin{aligned}
\triangle \widehat{\text{education}} &= 0.06 \triangle \text{income} \\
\triangle \widehat{\text{education}} &= 0.06 \times 100 \quad (\text{if } \triangle \text{income} = 100) \\
\triangle \widehat{\text{education}} &= 6
\end{aligned}
$$

- Answer: An increase of \$100 in per-capita income is associated with a predicted increase of \$6.00 in the average education expenditure

Summary

- **Today's Class:**
  - How to summarize the relationship between X and Y with a line: `lm()` and `geom_smooth()`.
  - How to interpret the two estimated coefficients: ($\widehat{\beta}_0$ and $\widehat{\beta}_1$) when outcome variable is non-binary.
  - How to make predictions with the fitted line:
    - Predict $\widehat{Y}$ based on X and predict.
    - 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 outcomes.
  - How to measure how well the model fits the data with R^2.

Questions?

See you in the next class!