Reviewing Regression (Pt. 1)

ECON 490

Taylor Mackay || Email: tmackay@fullerton.edu

Slides Overview

In these slides, we'll discuss:

- Characterizing relationships between variables
- A quick review of regression

Conditional Distributions

Last class, focused on the distribution of ***individual*** variables

- E.g., what proportion of people have X level of education?
- In general, this is known as an ***unconditional*** distribution

Conditional distributions tell us the distribution a variable given a specific value of another variable

Think about the sample ACS data set we had last week

- Variables including income, age, etc. for a sample of survey respondents
- We can use this data set to construct conditional distributions

Reference Slide – *ACS (American Community Survey) Sample Data*

Here's a snapshot of our sample data set:

state_name	*statefip*	*hhincome*	*sex*	*age*	*education*	*employed*	*food_stamp*	*renter*
california	6	133800	1	25	3	1	0	1
california	6	210000	1	40	3	1	0	0
california	6	157000	1	5	1	0	0	1
california	6	121600	2	62	2	1	0	1
california	6	84000	1	57	3	1	0	1

Definitions for selected key variables:

- `hhincome` is the total household income for the survey respondent over the past year
- `sex` is a binary variable equal to 1 if the respondent is male and 2 if female
- `education` is a factor variable with 3 levels (1 = Non-HS grad, 2 = HS grad, 3 = college grad)
- `employed` is a binary variable equal to 1 if someone is employed and 0 otherwise
- `food_stamp` is a binary variable equal to 1 if someone receives food stamps and 0 otherwise
- `renter` is a binary variable equal to 1 if someone rents their home and 0 if they own their home

A Simple Example to Start

What's the relationship between employment and food stamp recipiency?

- In other words, what's the conditional distribution of food stamp recipiency ***given*** whether or not someone is working?

How do we show this conditional distribution?

- Both variables are binary, so we can build on our approach from last week
- Use a bar plot to show the distribution of food stamp recipiency (either yes or no) given the values our employed variable can take (either yes or no)

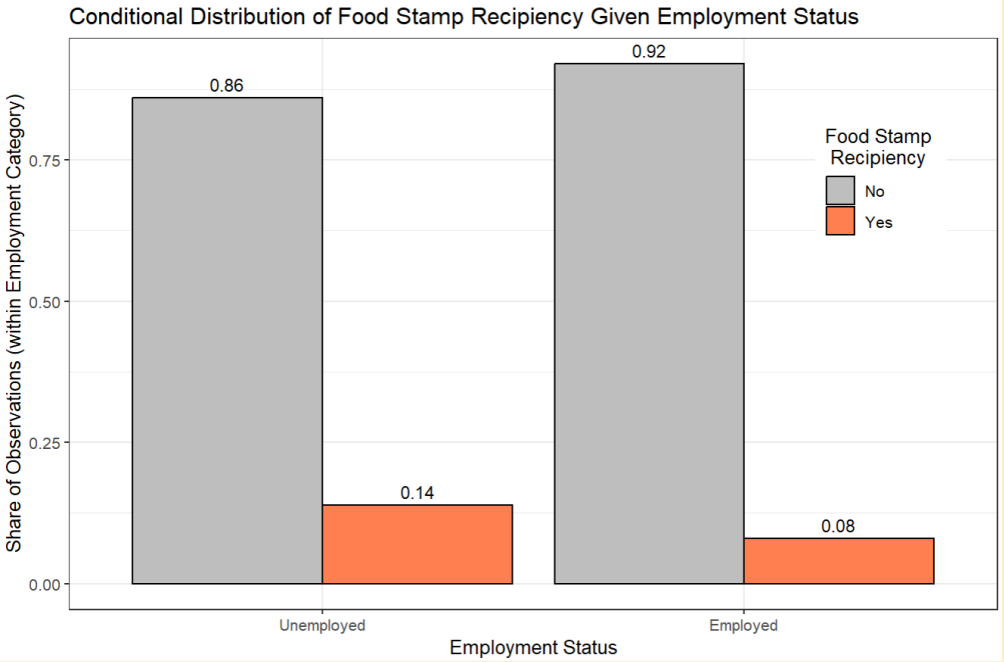
Graphing Conditional Distributions

Given two binary variables, its easy to plot conditional distributions

We're conditioning on employment

- X-axis is split across both values of employment (yes and no)
- Then we show food stamp recipiency rates across both groups

What if we want to condition on a continuous variable? More on this later…



Thinking Conditionally

Given two variables, we can also calculate ***conditional*** summary statistics

- Conceptually, this easier when we condition on discrete variables
- E.g., from last slide, average food stamp receipt for employed folks = 0.08

How do we condition on a ***continuous*** variable?

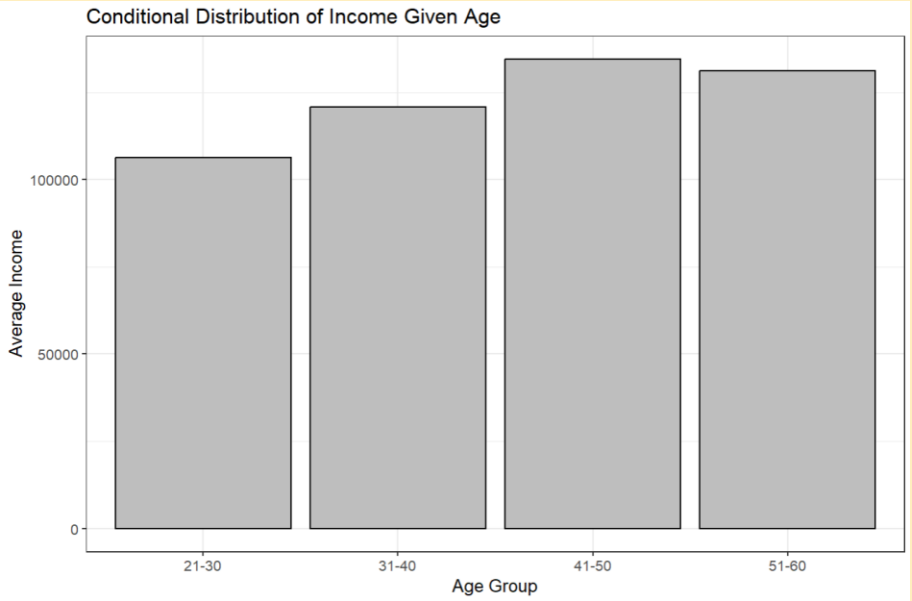
- E.g., what's average income at a particular age (or across the age distribution?)
- One option = use the "bins" approach we covered last week with histograms
- Create bins of age and calculate average income within those bins

Conditioning on a Continuous Variable

Same idea as with histograms – make our continuous age variable discrete

- Create 10-year bins of age
- Show average income in each bin

One question you might have – what if we ***don't*** make age discrete?



Relationships between Continuous Variables

This scatterplot shows age and income for a subsample of people in our data

- Now, age is continuous
- How to summarize this relationship?

One solution is to use a ***fitted line***

… but which what kind of fitted line?

Fitted Lines

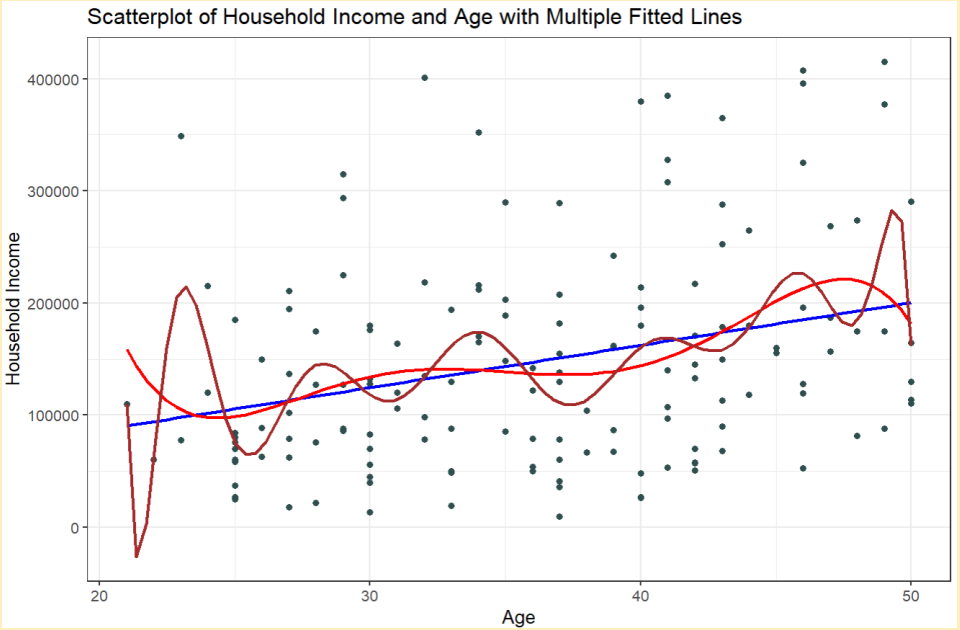
Using our data, we can construct lots of different fitted lines

- Given a particular age, each line calculates a mean value of income

What line is best?

- Not necessarily a single best answer
- Each line involves trade-offs

Let's focus on the blue (straight) line



Introducing Regression

We can generate the fitted lines from last slide using ***regression***

- Don't forget our original goal – describing conditional relationships

We're interested in how average income varies across ages

- Age is our explanatory (X) variable – it ***explains*** variation in income
- Income is our outcome (Y) variable – it's the ***outcome*** we'd like to explain

In broad terms, we're thinking about $Income = f(Age)$

- How do we pick f? Start with the simplest case from last slide, a straight line

Equation for the Slope of Our Fitted Line

We want to fit a straight line to characterize the income and age relationship

- Think back to slope-intercept formula from HS algebra → $y = mx + b$
- Putting this in terms of our data, we have $Income = mAge + b$

Instead of using m & b for the slope & intercept, economists would write this as:

$$Income = \beta_0 + \beta_1 Age$$

Different notation, same idea! Key question – how do we calculate β_0 and β_1?

OLS Regression

When we talk about fitted lines or regression in class, we'll (almost) always use this as shorthand for Ordinary Least Squares (OLS) regression

OLS is how we calculate β_0 and β_1 from last slide to draw our fitted line

- In other words, how we find m and b in our $y = mx + b$ point-slope equation
- In R, the `lm()` function calculates OLS regression for us

We'll talk about *how* OLS works later, but for now, let's look at *what* it tells us

Estimating our Regression Equation

Using R's `lm()` function, we can estimate our regression equation below:

$$Household\ Income = \beta_0 + \beta_1 Age$$

The output R gives us is the estimated values of our β_0 and β_1 coefficients

- These estimated coefficients are below
- We can use them to plot the blue fitted line we've seen previously

$$Household\ Income = \$10{,}735 + \$3{,}792 \times Age$$

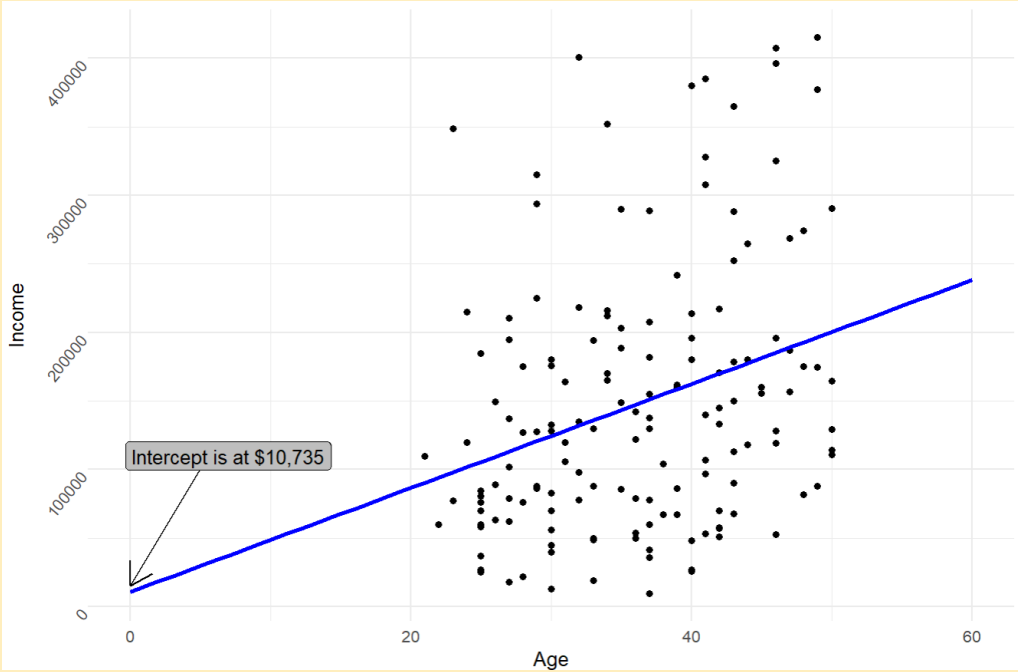
Plotting our Regression Equation

From the last slide, our equation was:

$$Household\ Income = \$10{,}735 + \$3{,}792 \times Age$$

Here, we've plotted this equation

- Data covers 20-to-50-year-olds
- The X-axis is extended so we can see the intercept



Interpreting our Regression Output

Here's the `lm()` output from last slide –

- There's a lot going on here! We'll unpack this over the next several classes

Key output for tonight:

- First line = our regression equation
- `Estimate` = our β coefficients
- `(Intercept)` = β_0
- age = β_1

```
lm(formula = hhincome ~ age, data = graph.data)

Residuals:
    Min      1Q  Median      3Q     Max
-141229  -68518  -21386   51575  268930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    10735      36505    0.29  0.76912
age             3792        989    3.83  0.00019 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 92800 on 148 degrees of freedom
Multiple R-squared:  0.0903,    Adjusted R-squared:  0.0842
F-statistic: 14.7 on 1 and 148 DF,  p-value: 0.000186
```

Using our Regression Output (Pt. 1)

What can we do with this output?

- Think back to our original goal – *describing a conditional relationship*
- In this case, describing the average level of income conditional on (or given) age

Our estimated regression let's us calculate average income conditional on age

- To do this, just plug an age value into our regression equation
- Suppose we want to know average income for 30-year-olds:

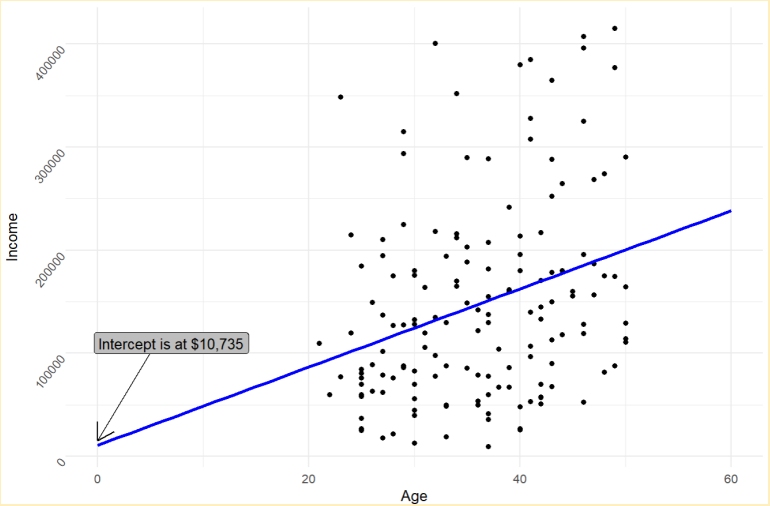
$$\begin{aligned} Average\ Household\ Income\ for\ 30-year-olds &= \$10{,}735 + \$3{,}792 \times 30 \\ &= \$10{,}735 + \$113{,}760 \\ &= \$124{,}495 \end{aligned}$$

Using our Regression Output (Pt. 1)

Suppose we repeated the process for every age from 0 to 60

- Plug each year into our equation
- Record average household income

If we plotted our average for every year, we'd wind up with the blue line



We know that at any given age, some folks earn more than other people…

- The benefit of regression (and our fitted line) is providing a convenient way of summarizing the income and age relationship
- In general, older folks tend to have higher household incomes (no surprises here!)

Using our Regression Output (Pt. 2)

In the last two slides, we talked about calculating conditional averages

- We can also calculate ***predicted values*** of income given age
- Different interpretation, same process (and output) as last two slides

What is the predicted value of income for someone 40 years old?

$$Household\ Income = \$10{,}735 + \$3{,}792 \times 40$$
$$= \$10{,}735 + \$151{,}680$$
$$= \$162{,}415$$

Interpreting Regression Coefficients

How does being one year older change predicted income? Let's see:

- Predicted income for a 40-year-old = $\$162{,}415$
- Predicted income for a 41-year-old = $\$10{,}735 + \$3{,}792 \times 41 = \$166{,}207$

What's the difference between our predictions?

- $\$166{,}207 - \$162{,}415 = \$3{,}792 =$ our estimated age coefficient β_1!

Importantly, this is true for ***any*** starting age – implies the following interpretation:

- *"A one-year increase in age is associated with a $3,792 increase in avg. income"*
- If you find interpreting regression coefficients tricky, try this approach!

Regression Residuals

We know income differs for all kinds of reasons besides age

- Implies our prediction is ***not*** going to be perfect for any given 40-year-old!
- ***Residuals*** are the difference between actual income and predicted income for each person in our regression data set

Residual = Actual Income – Predicted Income

From our data, we have records on the following 40-year-olds:

Person	Age	Actual Income	Predicted Income	Residual
1	40	$26,020	$162,415	-$136,395
2	40	$196,000	$162,415	$33,585
3	40	$48,000	$162,415	-$114,415

Regression Equations

At the start of these slides, “translated” point-slope $y = mx + b$ into $y = \beta_0 + \beta_1 X$

- Let’s us represent the slope of a fitted line…
- But this isn’t quite a “real” regression equation just yet!

Economists and statisticians write regression equations that look something like:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

Unpacking Regression Equations

Let's unpack the equation from last slide:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

There's several important components of this equation:

- Now, we have ***multiple*** explanatory variables: X_1 ***and*** X_2
- The i subscripts tell us how to identify each row of our regression data set
- We've added an ***error term*** u_i (more on this later)

A Couple Quick Examples

Example 1: $Hourly\ Wage_i = \beta_0 + \beta_1 Years\ of\ Education_i + \beta_2 Age_i + u_i$

- Our regression data set has earnings & demographic data for a sample of people, so we use i to index individuals in our data set
- Our outcome variable is $Hourly\ Wage$; explanatory variables are $Years\ of\ Education$ and Age

Example 2: $GDP_s = \beta_0 + \beta_1 Labor\ Force_s + \beta_2 Goverment\ Spending_s + u_s$

- In this example, we're now looking at ***state-level*** data – each row of our data is a state, so we'll index each row with s to make things clear
- Outcome variable is GDP; explanatory variables are $Labor\ Force$ and $Government\ Spending$

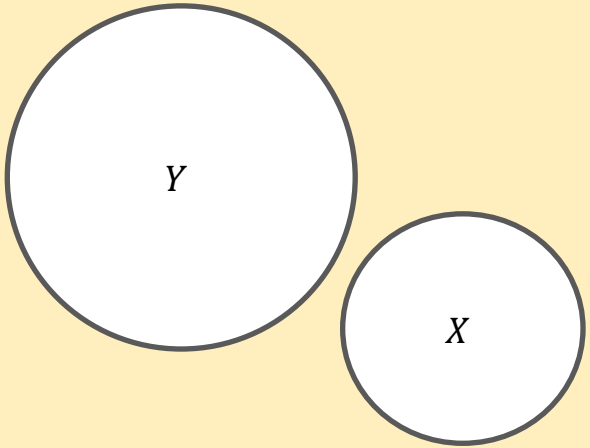
Representing Regression using Venn Diagrams

Up until now, we've talked about regression as a way of generating fitted lines

- This is how we'll generally think about regression
- The Venn diagram approach is another mental model that's useful

Suppose we're interested in the relationship between Y and X

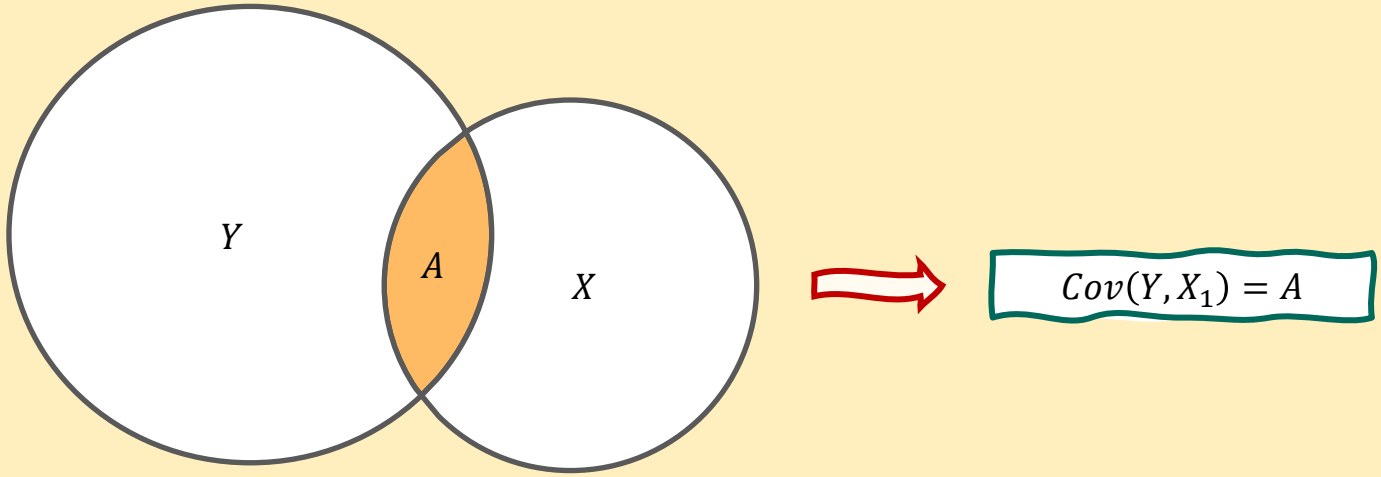
- Draw a circle for each variable
- The size of each circle indicates the ***variance*** of that variable



Visualizing Covariance

If Y and X were unrelated, then there'd be no overlap (like the last slide)

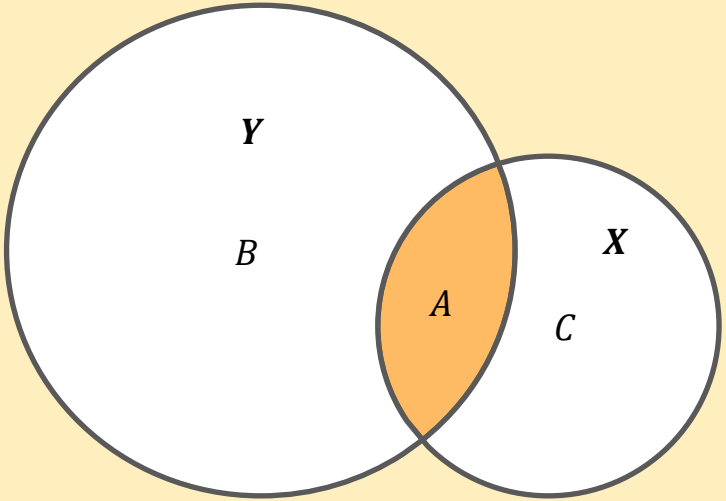
- If they're related, then they'll overlap
- The shaded region A below represents the ***covariance*** of Y and X



Regression Coefficients

From the last slide, we know the ***covariance*** of Y and X is A

- Suppose we ran the following regression: $Y = \alpha_0 + \alpha_1 X_1 + u$
- What is our estimated coefficient α_1?



Regression effect of X on Y is α_1:

$$\alpha_1 = \frac{Cov(Y,X)}{Var(X)} = \frac{A}{A+C}$$

Key Concepts for Quiz Next Week

Everything in slides is fair game, but the following concepts are important:

1. Defining conditional distributions
2. "Translating" the point-slope formula $y = mx + b$ (see slide 12)
3. Using regression output to calculate conditional averages and predicted values

Lastly, make sure to review the "Regression Residuals" slide

- You should be able to identify the origin of each column of data
- Our underlying data set has "Person," "Age," and "Actual Income" – we then use regression to (1) calculate "Predicted Income" and then (2) calculate "Residual"