Data Analysis in R

Interactions & Non-Linearity

Ken Stiller

13th January 2024

Syllabus: Data Analysis in R

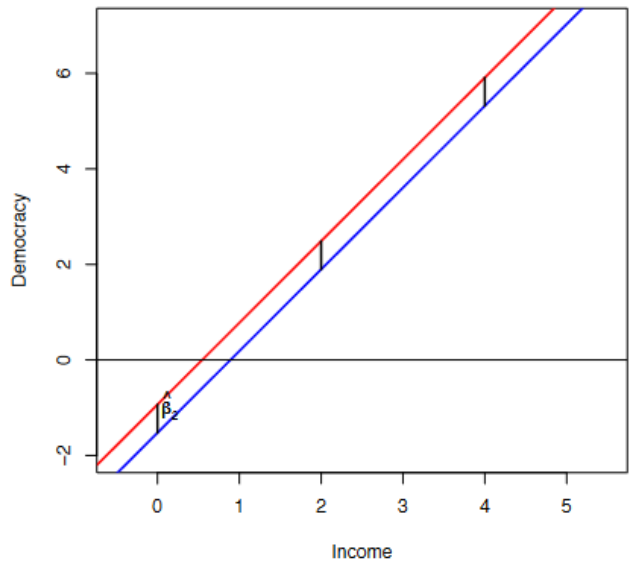
1. Introduction
2. Causality & Basics of Statistics
3. Sampling & Measurement
4. Prediction
5. Multivariate Regression
6. Probability & Uncertainty
7. Hypothesis Testing
8. Assumptions & Limits of OLS
9. **Interactions & Non-Linear Effects**

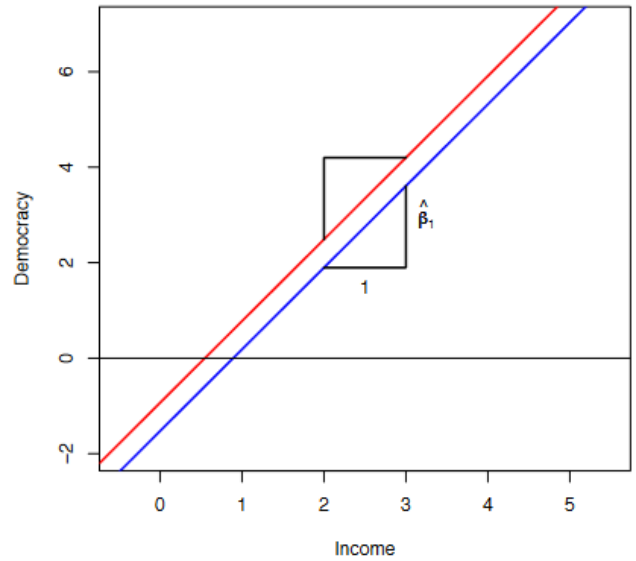
Overview

- Interactions
  - Logic of Interactions
  - Interpretation of regression tables, coefficient plots from actual papers
- Non-Linearity
  - Basic principle
  - Some examples
- Wrap up & Way ahead

Table of Contents

Interactions: Motivation





- So far, we have been assuming that effect of a treatment (independent variable) is the same for every unit
- But what if it isn't?
- Can you think of examples?

Interactions: Motivation II

Not surprisingly, many social phenomena are better understood if we take into account that the same event can affect people/units with distinct characteristics differently.

- Think of the effect of economic shocks on people with different SES
- This is the logic behind intersectionality
- The effect of policies may differ by political predisposition
- and so on...

Interactions: Example

- Are people who watch TV happier or less happy than people who do not watch TV?
- H_0: Watching TV is unrelated with level of happiness

- We use data from the European Social Survey. Happiness is measured by a 0 to 10 scale. TV is 1 if people watch TV more than 1 hour per day and 0 otherwise (*dummy variable*)
- We run the following simple model:

$$E[Hapiness|TV] = \beta_0 + \beta_1 * TV$$

Interactions: Example II

- What do we conclude?

```
Call:
lm(formula = happy ~ tv, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1398 -1.1398  0.2804  1.2804  3.2804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.13978    0.01283  556.30   <2e-16 ***
tv          -0.42020    0.01826  -23.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.127 on 54304 degrees of freedom
  (682 observations deleted due to missingness)
Multiple R-squared:  0.009657,  Adjusted R-squared:  0.009639
F-statistic: 529.5 on 1 and 54304 DF,  p-value: < 2.2e-16
```

Interactions: Example III

- According to the regression table, watching TV is associated with lower rates of happiness. This association is statistically significant and substantively large (0.4 in a 10-point scale; around 6% of baseline)
- This finding summarises the average relationship between TV and happiness. But does this negative association hold for all groups of people?

Imagine you have reason to believe that this is different for widowed people

- Let's check whether for this particular group, widowed people, the relationship between TV and happiness is different
- To do so, we will estimate the same regression but now only for the subset of people who are widowed

Interactions: Example III

- According to the regression table, watching TV is associated with lower rates of happiness. This association is statistically significant and substantively large (0.4 in a 10-point scale; around 6% of baseline)
- This finding summarises the average relationship between TV and happiness. But does this negative association hold for all groups of people?

Imagine you have reason to believe that this is different for widowed people

- Let's check whether for this particular group, widowed people, the relationship between TV and happiness is different
- To do so, we will estimate the same regression but now only for the subset of people who are widowed

Interactions: Example IV

```
> tvwid<-tv[tv$widow==1, ]
>
> summary(lm(happy ~ tv,data=tvwid))

Call:
lm(formula = happy ~ tv, data = tvwid)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9329 -0.9329  0.0671  2.0671  4.1517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.84832    0.05416 107.987   <2e-16 ***
tv           0.08456    0.06773   1.249    0.212
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.417 on 5520 degrees of freedom
```

Why We Need Interactions

- We face a problem: our naive model doesn't account for this difference
- Recall: In the model $E[Hapiness|TV] = \beta_0 + \beta_1 * TV$, β_1 is the same for **all** units
- We know now that we should actually relax this assumption for at least one group, widowed people
- We want a single model that accounts for this difference in effects

- We can specify such model by including interaction terms
- The model distinguishes units by there status of being widowed or not. The latter variable is called a moderator

Why We Need Interactions

- We face a problem: our naive model doesn't account for this difference
- Recall: In the model $E[Hapiness|TV] = \beta_0 + \beta_1 * TV$, β_1 is the same for **all** units
- We know now that we should actually relax this assumption for at least one group, widowed people
- We want a single model that accounts for this difference in effects

- We can specify such model by including interaction terms
- The model distinguishes units by there status of being widowed or not. The latter variable is called a moderator

Interactions: The Set-Up

Two independent variables, X_1 (treatment):

$$TV_i = \begin{cases} 1 & \text{if watch TV more than one hour per day} \\ 0 & \text{if not watch TV more than one hour per day} \end{cases}$$

and X_2 (moderator):

$$\text{Widowed} = \begin{cases} 1 & \text{The person is widowed} \\ 0 & \text{anything else} \end{cases}$$

- Until now, the only thing we did was to control for X_2 in order to isolate its effect on Y from the effect of X_1 on Y
- Here, our problem is not that β_1 is biased because of a confounder, but that it can be different for the widowed subset.
- To see the problem, let's regress happiness on TV consumption and marital status:
  $Happiness_i = \beta_0 + \beta_1 * X_{TV} + \beta_2 * X_{Widowed} + \epsilon_i$

Interactions: Results I

What is the average level of happiness in each category?

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{Widowed} + \epsilon_i$$

	$X_{\text{Wid}} = 1$	$X_{\text{Wid}} = 0$
$X_{TV} = 1$	?	?
$X_{TV} = 0$	?	?

```
> summary(lm(happy ~ tv + widow,data=tv))

Call:
lm(formula = happy ~ tv + widow, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2187 -1.2187  0.1371  1.2259  4.2259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.21868    0.01287  561.05   <2e-16 ***
tv          -0.35579    0.01813  -19.62   <2e-16 ***
widow       -1.08878    0.02999  -36.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.102 on 54303 degrees of freedom
  (682 observations deleted due to missingness)
Multiple R-squared:  0.03312,   Adjusted R-squared:  0.03309
```

Interactions: Results II

What is the average level of happiness in each category?

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{Widowed} + \epsilon_i$$

	$X_{\text{Wid}} = 1$	$X_{\text{Wid}} = 0$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{\text{WID}}$	$\beta_0 + \beta_{TV}$
$X_{TV} = 0$	$\beta_0 + \beta_{\text{WID}}$	β_0

```
> summary(lm(happy ~ tv + widow,data=tv))

Call:
lm(formula = happy ~ tv + widow, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2187 -1.2187  0.1371  1.2259  4.2259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.21868    0.01287  561.05   <2e-16 ***
tv          -0.35579    0.01813  -19.62   <2e-16 ***
widow       -1.08878    0.02999  -36.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.102 on 54303 degrees of freedom
  (682 observations deleted due to missingness)
Multiple R-squared:  0.03312,   Adjusted R-squared:  0.03309
```

Interactions: Results III

What is the average level of happiness in each category?

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{Widowed} + \epsilon_i$$

	$X_{\text{Wid}} = 1$	$X_{\text{Wid}} = 0$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{\text{WID}} = 5.77$	$\beta_0 + \beta_{TV} = 6.87$
$X_{TV} = 0$	$\beta_0 + \beta_{\text{WID}} = 6.12$	$\beta_0 = 7.22$

```
> summary(lm(happy ~ tv + widow,data=tv))

Call:
lm(formula = happy ~ tv + widow, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2187 -1.2187  0.1371  1.2259  4.2259

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.21868    0.01287  561.05   <2e-16 ***
tv          -0.35579    0.01813  -19.62   <2e-16 ***
widow       -1.08878    0.02999  -36.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.102 on 54303 degrees of freedom
  (682 observations deleted due to missingness)
Multiple R-squared:  0.03312,   Adjusted R-squared:  0.03309
```

The Problem: No Group Differentiation

What do the β tell us?

- What is the predicted difference in average level of happiness between widowed respondents who do [not] watch TV and non-widowed respondents who [do not] watch TV?
- Note that the effects are separate and uniform

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{WID} = 5.77$	$\beta_0 + \beta_{TV} = = 6.87$	β_{WID}
$X_{TV} = 0$	$\beta_0 + \beta_{WID} = 6.12$	$\beta_0 = 7.22$	β_{WID}
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	β_{TV}	β_{TV}	

Adding an Interaction Term

Now, consider our two independent variables, X_{TV} (treatment):

$$TV_i = \begin{cases} 1 & \text{if watch TV more than one hour per day} \\ 0 & \text{if not watch TV more than one hour per day} \end{cases}$$

and X_{WID} (moderator):

$$\text{Widowed} = \begin{cases} 1 & \text{The person is widowed} \\ 0 & \text{anything else} \end{cases}$$

and an interaction term:

$$X_{TV \times WID} = X_{TV} \times X_{WID} = \begin{cases} 1 & \text{if unit } i \text{ is widowed and watches TV} \\ 0 & \text{otherwise - do you see why?} \end{cases}$$

$$\text{Y}=\beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + \beta_3 X_{TV} X_{WID} + u_i$$

What Do our Coefficients Tell Now?

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + \beta_3 X_{TV} X_{WID} + u_i$$

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{WID} + \beta_{TV \times WID}$	$\beta_0 + \beta_{TV}$	$\beta_{WID} + \beta_{TV \times WID}$
$X_{TV} = 0$	$\beta_0 + \beta_{WID}$	β_0	β_{WID}
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	$\beta_{TV} + \beta_{TV \times WID}$	β_{TV}	

```
> summary(lm(happy ~ tv+widow+tv*widow,data=tv))

Call:
lm(formula = happy ~ tv + widow + tv * widow, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2407 -1.2407  0.1612  1.1612  4.1517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.24067    0.01316  550.19  < 2e-16 ***
tv          -0.40185    0.01904  -21.10  < 2e-16 ***
widow       -1.39235    0.04889  -28.48  < 2e-16 ***
tv:widow     0.48641    0.06189    7.86 3.91e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.101 on 54302 degrees of freedom
  (682 observations deleted due to missingness)
```

Comparing the Two Models

- Without interaction term:

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + u_i$$

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{WID}$	$\beta_0 + \beta_{TV}$	β_{WID}
$X_{TV} = 0$	$\beta_0 + \beta_{WID}$	β_0	β_{WID}
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	β_{TV}	β_{TV}	

- With interaction term:

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + \beta_3 X_{TV} X_{WID} + u_i$$

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	$\beta_0 + \beta_{TV} + \beta_{WID} + \beta_{TV \times WID}$	$\beta_0 + \beta_{TV}$	$\beta_{WID} + \beta_{TV \times WID}$
$X_{TV} = 0$	$\beta_0 + \beta_{WID}$	β_0	β_{WID}
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	$\beta_{TV} + \beta_{TV \times WID}$	β_{TV}	

Output With Interaction Term

```
> summary(lm(happy ~ tv+widow+tv*widow,data=tv))

Call:
lm(formula = happy ~ tv + widow + tv * widow, data = tv)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2407 -1.2407  0.1612  1.1612  4.1517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.24067    0.01316  550.19  < 2e-16 ***
tv          -0.40185    0.01904  -21.10  < 2e-16 ***
widow       -1.39235    0.04889  -28.48  < 2e-16 ***
tv:widow     0.48641    0.06189    7.86 3.91e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.101 on 54302 degrees of freedom
  (682 observations deleted due to missingness)
Multiple R-squared:  0.03422.   Adjusted R-squared:  0.03417
```

Comparing the Two Models II

- Without interaction term:

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + u_i$$

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	5.77	6.87	-1.10
$X_{TV} = 0$	6.12	7.22	-1.10
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	-0.35	-0.35	

- With interaction term:

$$Y = \beta_0 + \beta_1 X_{TV} + \beta_2 X_{WID} + \beta_3 X_{TV} X_{WID} + u_i$$

	$X_{\text{WID}} = 1$	$X_{\text{WID}} = 0$	$E[Y \mid X_{WID} = 1, X_{TV}] - E[Y \mid X_{WID} = 0, X_{TV}]$
$X_{TV} = 1$	5.933	6.839	-0.92
$X_{TV} = 0$	5.848	7.241	-1.40
$E[Y \mid X_{TV} = 1, X_{WID}] - E[Y \mid X_{TV} = 0, X_{WID}]$	0.085	-0.402	

Interpreting the Interaction Term

- $\beta_{WID \times TV}$ is large and significant, indicating that there is an important interaction between the two variables
- Statistical significance here tests the null that there is no difference in association across the two-groups (widowed and non-widowed)
- The expectation for people who are widowed and watch TV is he sum of intercept and all three coefficients: watch TV, widowed, and interaction thereof
- Important: Software and math do not know which one is the treatment and which one is the moderator: interaction can be read both ways and it is up to you to interpret appropriately

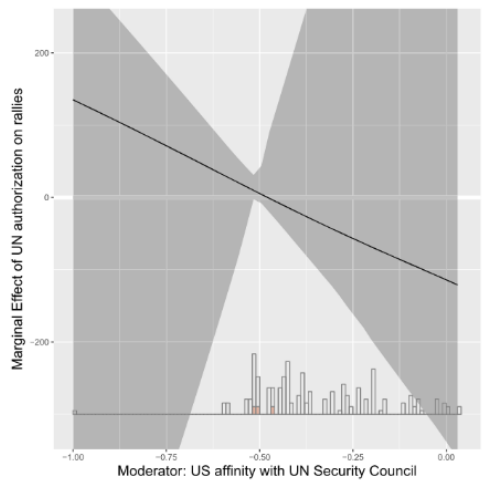
Interaction With Continuous Moderators

- You can no longer nicely build the table with all the possible values like we did for the dummy variables before (as this would be an infinite number of values)
- Adding up the coefficients will give you the predicted outcome for someone with a value of 1 in the treatment and 1 in the moderator (recall: coefficients are about one-unit changes)
- How much this makes sense depends on the case at hand: always think of the substantive interpretation of the results
- What you *can* do is to calculate how the effect of the treatment looks like at different values of the moderator
- **Don't be mislead by statistical significance! Usually, plotting the interaction effect gives you a good idea of its substantive meaning**

Words of Caution

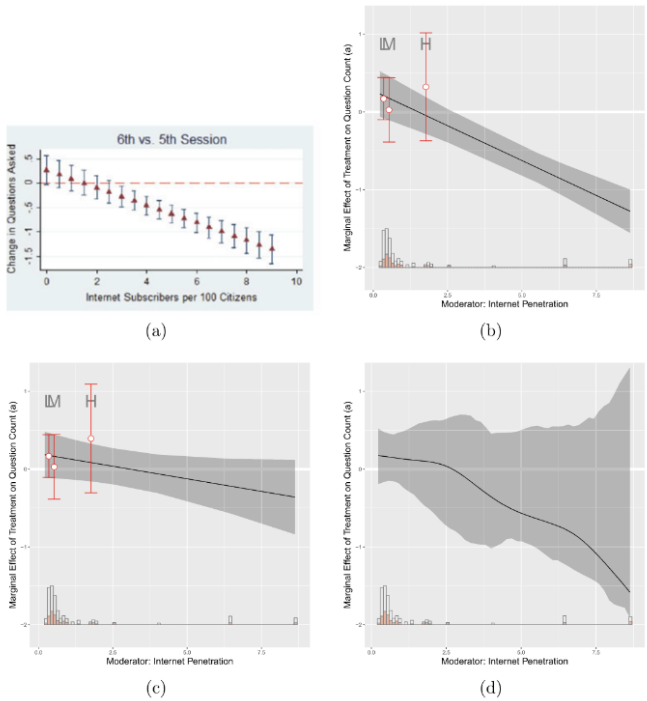
- Make sure there is common support. To compute the effect of treatment on outcome across values of the moderator (X_m), you need
  1. a sufficient number of observations whose values of the moderator are (close to) X_m
  2. variation in the treatment at X_m

- What you don't want to see (from Hainmueller, Mummolo and Xu, PA 2018):



Words of Caution II

- Be careful with extrapolations (and interpolations). You *can* calculate effects for any value of the moderator but always bear in mind what you actually get in the data
- What you don't want to see (from Hainmueller, Mummolo and Xu, PA 2018):



Words of Caution III

- Recall that the OLS estimator is agnostic to the role of independent variables
- Ideally, make sure the moderator is pre-treatment
  - Interaction models include the moderator as another independent variable
  - Thus, if the moderator is post-treatment you may reintroduce selection bias
- Interactions require quite some statistical power
- Make sure your sample size is large enough

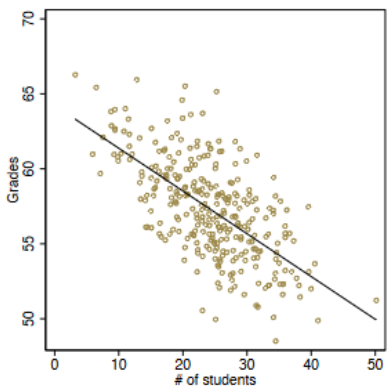
Linearity

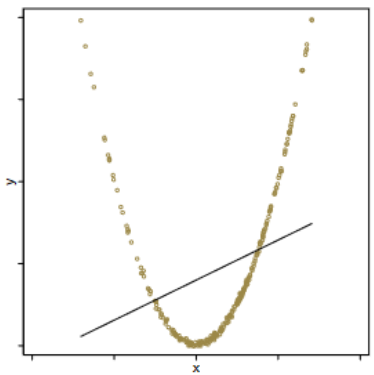
- Under OLS, The population regression model we have used is linear in the parameters.
- Y is related to X and the unobserved error ϵ as:

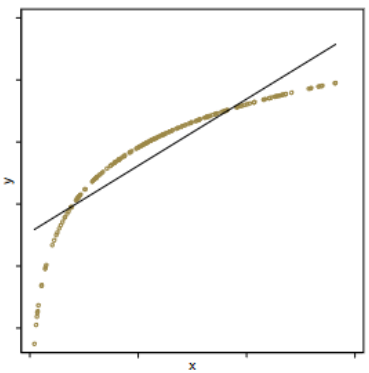
$$Y = \beta_0 + \beta_1 \times X + \epsilon$$

where β_0 & β_1 represent the fixed and unknown regression parameters

- Yet, there are (many) cases where the linear model has a poor fit:







- Luckily, we can model these relationships in a more flexible way: $Y = \beta_0 + \beta_1 f(X) + \epsilon$

Non-Linearity

- Polynomial terms can be used to model non-linearities.
- For example, we can estimate a model with a linear and a quadratic term:

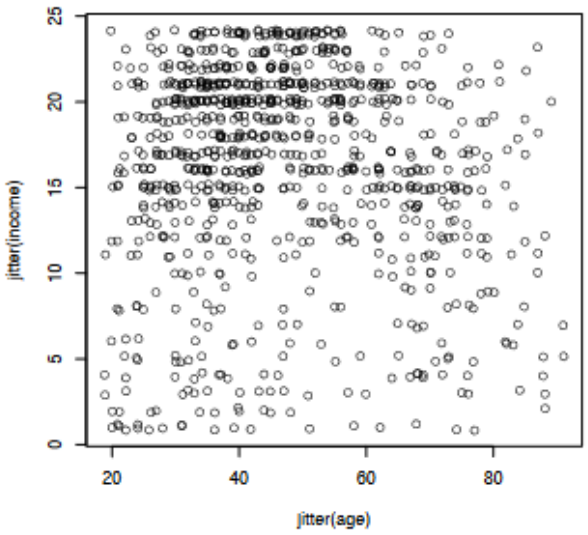
$$Y = \beta_0 + \beta_1 \times X + \beta_2 X^2 + \epsilon$$

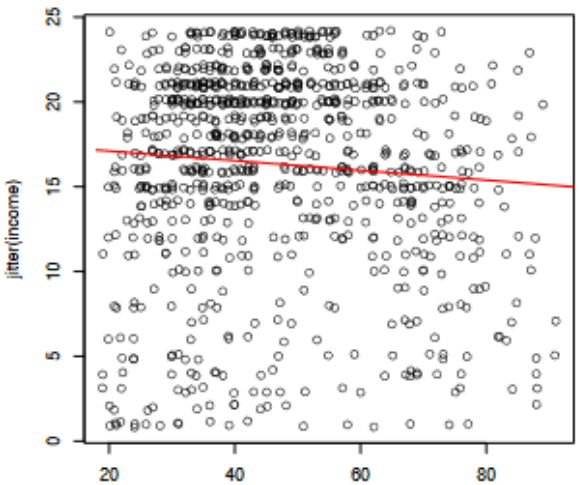
- This is also known as a second-order polynomial in X
- We could also use a third-order polynomial if need be:

$$Y = \beta_0 + \beta_1 \times X + \beta_2 X^2 + \beta_3 X^3 + \mu$$

- Don't include the higher polynomial without including also lower polynomials, e.g. never include only X^2 without also including X (the linear/scalar term)

Example: Income and Age



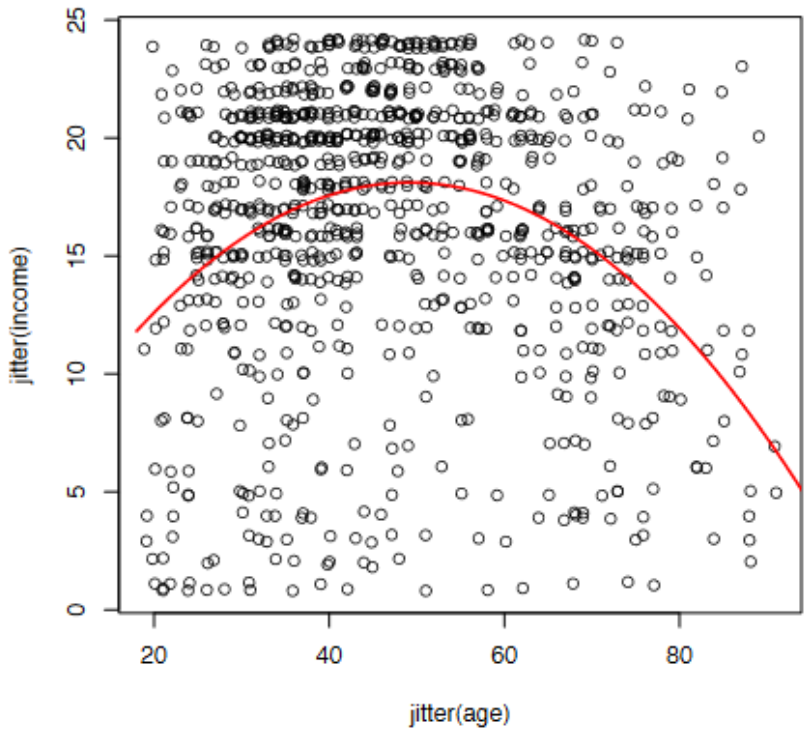


- Y = income, X = age
- Let's try a linear specification first
- $Y = \beta_0 + \beta_1 X_1 + \epsilon$

- When you see such a pattern, always think of adding the quadratic term, hence fitting a second-order polynomial:

$$Inc. = \beta_0 + \beta_1 Age + \beta_2 Age^2 + \epsilon$$

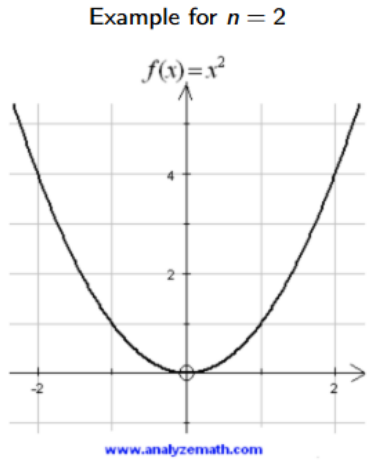
Example: Income and Age II

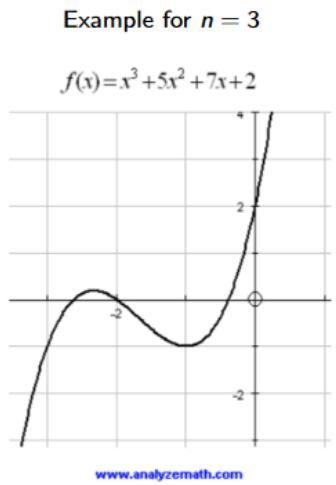


- Much better fit now
- β_2 is the marginal effect of age
- Now, the conditional marginal effect of age depends on the level of age: $\beta_1 + \beta_2 X_1$
- Here, the effect of age is first increasing and then decreasing with income
- When $\beta_2 > 0$, we get a U-shaped curve (there is no maximum, there is a minimum)
- When $\beta_2 < 0$, we get an inverted U-shaped curve (there is no minimum, there is a maximum, as in the figure)

Higher-Order Polynomials

- Let f be the polynomial function $f(X) = a_n X^n + ... + a_1 X + a_0$
- When n is odd, the graph goes in a different direction when X goes left than when X goes right
- When n is even, the graph goes in the same direction when X goes left and right
- A polynomial function f has $n - 1$ inflexion points
- Again, if you have n polynomials, you need to include all lower-order polynomials up to n





The Downside of Polynomials

- Polynomials are a powerful, flexible way of dealing with non-linearities
- The major downside is: we mostly want results that we can extrapolate to a general context
- There is some trade-off between fitting your sample and being able to extrapolate to different samples
- Why though?

Why Do We Analyse Data?

- What we have been trying to do is to have best guesses at the value of a variable in the population based on a sample
- In the simplest case, we draw upon descriptive statistics (measures of central tendency and spread)
- If we know more variables, we can look at associations: univariate or multivariate regressions
- **Our aim is to isolate causal effects**
- Controlling for confounders aims to help us achieve this
- We always care about size of effects **and** their statistical significance

- If you suspect your relation is not the same for all units, you can add interactions
- If you suspect your relation is not linear, you can add polynomials (but be careful that this can still be extrapolated)

The Usual Process

1. Start with theory
   - What is out there?
   - Find your niche, try to contribute to debate
   - Then formulate hypotheses

2. Move to data collection
   - What is your population of interest?
   - Can you get a (random) sample?
   - If not, how much can you extrapolate?
3. Think of the right model (specification)
   - Is your treatment randomly assigned?
   - Otherwise, what confounders should you account for?
   - Can you measure them?
   - Need to re-specify model?
4. Interpret results: size *and* significance

Why Should You Care?

- Quantitative literature is everywhere: even if you don't want to do it, you will consume it
- Beyond academic literature, quantitative thinking is ubiquitous. In public discourse, people make mistakes we should be careful of:
  - Assuming $P(A|B) = P(B|A)$ (e.g., vaccines, police shootings, etc.)
  - Not thinking of possible confounders
  - Extrapolating from non-random samples

Why Should You Care? II

- Quantitative and qualitative analyses can [should] complement each other:
  - Qual. analyses help understand mechanism of quant. analyses. Example: Abdelgadir and Fouka 2020, APSR

American Political Science Review (2020) 114, 3, 707–723
doi:10.1017/S0003055420000106

Political Secularism and Muslim Integration in the West: Assessing the Effects of the French Headscarf Ban

AALA ABDELGADIR *Stanford University*
VASILIKI FOUKA *Stanford University*

In response to rising immigration flows and the fear of Islamic radicalization, several Western countries have enacted policies to restrict religious expression and emphasize secularism and Western values. Despite intense public debate, there is little systematic evidence on how such policies influence the behavior of the religious minorities they target. In this paper, we use rich quantitative and qualitative data to evaluate the effects of the 2004 French headscarf ban on the socioeconomic integration of French Muslim women. We find that the law reduces the secondary educational attainment of Muslim girls and affects their trajectory in the labor market and family composition in the long run. We provide evidence that the ban operates through increased perceptions of discrimination and that it strengthens both national and religious identities.

  - Qualitative studies can *help generate hypotheses* to test with a larger N
  - Looking at a distribution of cases can help with case selection for case studies / small-N studies

Beyond OLS: Modelling for Different Types of DV

What if my dependent variable isn't continuous?

- Logistic regression for Binary Dependent Variables
- Multinomial regressions for Categorical Dependent Variables
- Ordered logit for Ordinal Dependent Variables

  - Agresti (2018) *Statistical Methods for the Social Sciences*, Chaps. 14-15.
  - Agresti (2013) *Categorical Data Analysis.*
  - Gelman and Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Part 1).

Beyond OLS: Modelling Different Types of Data

What if my observations are not independent from each other?

- Multilevel Modelling for nested data (units are 'grouped' and we want to estimate unit- and group-level effects)
  - Gelman and Hill (2007) Data Analysis Using Regression and Multilevel/Hierarchical Models (Part 2)

- Panel Data and Event History analysis for times-series (same units are observed at different times)
  - Woolridge (2013) *Introductory Econometrics: A Modern Approach* (Chapters 10-14), available here
  - Kleinbaum and Klein (2011) *Survival Analysis: A Self-Learning Text.*

Beyond OLS: Other Applications

- Measurement (Classification, PCA, Factor Analysis, IRT)
  - Lauderdale (forthcoming) *Pragmatic Social Measurement*, available here
  - Kabacoff (2021) *R in Action*, Chaps. 14-17, available here

- Statistical Learning (aka 'Machine Learning')
  - James, Witten, Hastie and Tibshirani (2021) *An Introduction to Statistical Learning*, available here
  - Lantz (2013) *Machine Learning with R*, available here
- Qualitative Comparative Analysis (QCA)
  - Schneider, Thomann and Oana (2021) *Qualitative Comparative Analysis Using R: A Beginner's Guide*

Beyond Base R: Making the Most out of the Software

- Data Wrangling, Working with String Variables, Creating Functions, and more...
  - Chris Hanretty's short course 'ConverRt to R', available here
  - Wickham and Grolemund (2016) *R for Data Science*, the "Bible" of `tidyverse` users, available here

- Data Visualisation
  - Healy (2019) *Data Visualization: A Practical Introduction*

- Spatial Data and Maps in R
  - Lovelace Nowosad and Muenchow (2021) *Geocomputation with R*, available here

- RMarkdown
  - Xie, Allaire and Grolemund (2021) *R Markdown: The Definitive Guide*, available here

Thank you and all the best!

Feel free to get in touch about anything you think I can help with