Quantitative Methods in Social Research

Burak Sonmez

UCL

Feb 20, 2026 (UPF)

Day 3 – Outline

1. Understanding Interaction Effects
2. Non-linear Models and Their Interactions
3. Analysing Nested Data: Country Effects
4. Survey Weighting

Day 3 – Outline

1. Understanding Interaction Effects
2. Non-linear Models and Their Interactions
3. Analysing Nested Data: Country Effects
4. Survey Weighting

Interaction terms

- In a regression model is it assumed that each explanatory variable is related to the outcome variable independent of all others.

Interaction terms

- In a regression model is it assumed that each explanatory variable is related to the outcome variable independent of all others.
- Sometimes the effect of an explanatory variable depends on the level of another explanatory variable.

Interaction terms

- In a regression model is it assumed that each explanatory variable is related to the outcome variable independent of all others.
- Sometimes the effect of an explanatory variable depends on the level of another explanatory variable.
- If this occurs, it is called an interaction effect and we should include it in the model.

Interaction terms

- In a regression model is it assumed that each explanatory variable is related to the outcome variable independent of all others.
- Sometimes the effect of an explanatory variable depends on the level of another explanatory variable.
- If this occurs, it is called an interaction effect and we should include it in the model.

OLS with an interaction

- Consider the following regression equation:
- We can use differentiation to find the marginal effects of X and Z on Y

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + u$$

- Marginal effect of X: $\delta_y/\delta_x = \beta_1 + \beta_3 Z$
- Marginal effect of Z: $\delta_y/\delta_z = \beta_1 + \beta_3 X$

- Interpretation: a one-unit change in X is associated with a $\beta_1 + \beta_3 Z$ change in Y

Overfitting issues

- When you have large number of independent variables, it is not recommended to add/investigate all possible interaction terms in your model.

Overfitting issues

- When you have large number of independent variables, it is not recommended to add/investigate all possible interaction terms in your model.
- You should use theoretical reasoning for their inclusion or those suggested by descriptive analysis.

Overfitting issues

- When you have large number of independent variables, it is not recommended to add/investigate all possible interaction terms in your model.
- You should use theoretical reasoning for their inclusion or those suggested by descriptive analysis.
- Keep the main effects even if they become insignificant.

Overfitting issues

- When you have large number of independent variables, it is not recommended to add/investigate all possible interaction terms in your model.
- You should use theoretical reasoning for their inclusion or those suggested by descriptive analysis.
- Keep the main effects even if they become insignificant.

Marginal effects and OLS

In a model with no interactions:

- The coefficients directly govern the relationship between Y and X
- The marginal effect is constant and equal to a single coefficient.
- Changes in the coefficient are the same thing as changes in the marginal effect

Marginal effects and OLS

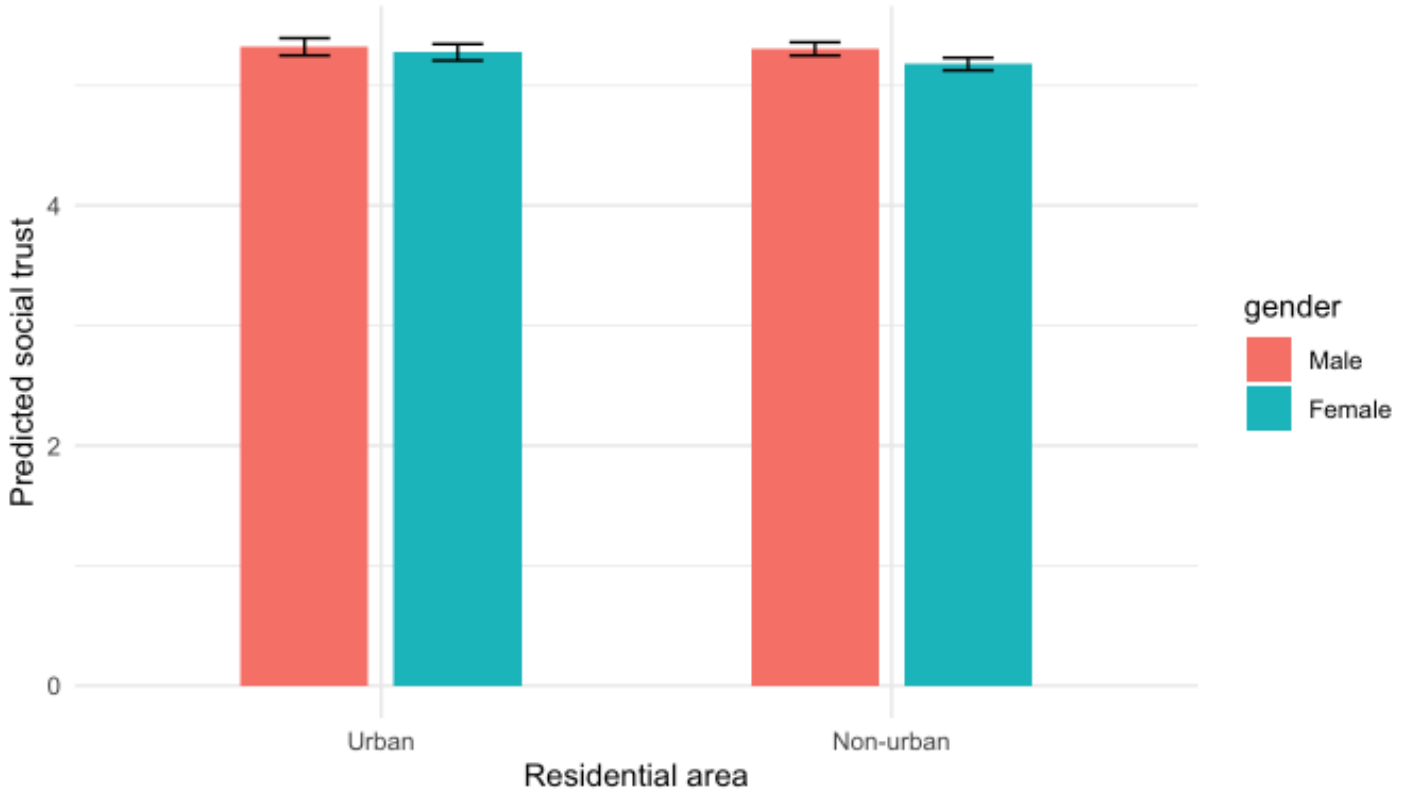
In a model with no interactions:

- The coefficients directly govern the relationship between Y and X
- The marginal effect is constant and equal to a single coefficient.
- Changes in the coefficient are the same thing as changes in the marginal effect

In a model with interactions:

- The coefficients directly govern the relationship between Y and X
- The marginal effect is not constant, but changes in the coeffcients directly translate into changes in the marginal effects.
- The regression results still contain all the information needed to study the relationship between Y and X

Is the relationship between generalised trust and residential area similar by respondent's gender?



Day 3 – Outline

1. Understanding Interaction Effects
2. Non-linear Models and Their Interactions
3. Analysing Nested Data: Country Effects
4. Survey Weighting

Moving Beyond Linear Models

Binary Outcomes

When our dependent variable is binary (0/1), we need different approaches than OLS

Our Example:

- Outcome: `news_regular` = 1 if respondent reads newspapers ≥ 3 days/week, 0 otherwise
- Predictors: age (`agea`), gender (`gndr`), country (`cntry`), education (`eduyrs`)

Three specifications we'll cover:

1. Linear Probability Model (LPM)
2. Logistic Regression
3. Interactions in logit models

Linear Probability Model (LPM)

Specification:

$$Y_i = X_i\beta + \varepsilon_i, \quad Y_i \in \{0, 1\}$$

Ordinary least squares with a binary outcome: $E[Y_i|X_i] = X_i\beta$

Pros:

- Coefficients are immediate probability changes (Δp) per unit of X
- Easy to interpret
- Simple to add fixed effects
- Fast estimation

LPM: Limitations

Cons:

- Predicted values can leave [0,1] range
- Errors are heteroskedastic by construction
- Marginal effects assumed constant even when baseline risk is near 0 or 1

When is LPM "safe enough"?

- Middle-range probabilities (e.g., 0.2–0.8)
- Modest leverage points
- Goal is fast descriptive decomposition or fixed-effects absorption
- **Always use robust standard errors**

LPM Example

```
# Linear Probability Model
lpm <- lm(
  news_regular ~ agea + gender + eduyrs + country,
  data = ess
)

# With robust standard errors
lpm_robust <- coeftest(lpm, vcov = vcovHC(lpm, type = "HC1")
    )
tidy(lpm_robust)
```

Interpretation:

- Each coefficient = change in probability (0–100 percentage points)
- Example: $\beta_{\text{agea}} = 0.003$ means each additional year of age increases probability of regular news reading by 0.3 percentage points

Logistic Regression

Specification:

$$p_i = Pr(Y_i = 1|X_i) = \text{logit}^{-1}(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

Advantages over LPM:

- Predictions always in [0,1]
- S-shaped curve matches reality: effects diminish near boundaries
- Better statistical properties

Trade-off:

- Coefficients are log-odds, not probabilities
- Need to compute marginal effects for interpretation

Logit Example

```
# Logistic regression
logit1 <- glm(
  news_regular ~ agea + gender + eduyrs + country,
  data = ess,
  family = binomial()
)

tidy(logit1)
```

Interpretation of coefficients:

- $\beta = 0.5$ means a one-unit increase in X multiplies the odds by $e^{0.5} \approx 1.65$
- Positive β = higher probability; negative β = lower probability
- Magnitude hard to interpret directly — use marginal effects

Marginal Effects in Logit

The key formula:

$$\frac{\partial p_i}{\partial x_{ik}} = p_i(1 - p_i)\beta_k$$

Key insight: Effects vary with baseline probability!

- When $p_i = 0.5$: Maximum effect $= 0.25\beta_k$
- When $p_i \to 0$ or $p_i \to 1$: Effect approaches 0
- Same β has different impact depending on where you are on the curve

Example

If $\beta_{\text{age}} = 0.02$ and baseline $p = 0.5$, then marginal effect $=$ $0.5 \times 0.5 \times 0.02 = 0.005$ (0.5 percentage points per year)

Computing Marginal Effects

```
# Average marginal effects
library(margins)

# Marginal effects at the mean
logit1_mfx <- margins(logit1)
summary(logit1_mfx)

# Or marginal effects at representative values
margins(logit1, at = list(agea = c(30, 50, 70)))
```

Recommendation:

- Report marginal effects, not raw coefficients
- Consider effects at different baseline probabilities
- Average marginal effects (AME) are often most useful

Interactions in Logit: The Challenge

For model: $\eta_i = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$ with $p_i = \text{logit}^{-1}(\eta_i)$

The cross-partial effect:

$$\frac{\partial^2 p_i}{\partial x \partial z} = p_i(1 - p_i)(1 - 2p_i)\beta_1\beta_2 + p_i(1 - p_i)\beta_3$$

Important!

- Sign can vary with p_i
- The interaction effect is NOT just β_3
- Must be evaluated at specific values
- Visualization is crucial

Binary × Binary Interaction

Example: Gender × Country

```
# Logit with interaction
logit2 <- glm(
  news_regular ~ gender * country + agea,
  data = ess,
  family = binomial()
)

# Predicted probabilities by group
pred_probs <- ggpredict(logit2, terms = c("country", "gender
    "))
plot(pred_probs)
```

Interpretation:

- If male–female confidence intervals overlap within countries: gender differences are modest
- Compare across countries to see where the gap is largest

Continuous × Binary Interaction

Example: Age × Gender

```
# Age[U+FFFD]gender interaction
logit3 <- glm(
  news_regular ~ agea * gender + eduyrs + country,
  data = ess,
  family = binomial()
)

# Plot predicted probabilities
pred_age_gender <- ggpredict(
  logit3,
  terms = c("agea [20:80]", "gender")
)
plot(pred_age_gender)
```

Look for:

- Do the lines cross?
- Are they parallel (no interaction) or diverging (interaction)?

LPM vs Logit: Decision Guide

Use LPM when:

- Quick descriptive analysis
- Need many fixed effects
- Probabilities mostly 0.2–0.8
- Want simple interpretation
- Comparison across subgroups

Use Logit when:

- Probabilities near 0 or 1
- Publication standard
- Correct functional form matters
- Building predictive model
- Theory suggests S-curve

Pragmatic Approach

- Fit both, compare substantive conclusions
- If results similar: LPM is easier
- If results differ: Logit is safer
- Always report which you used and why

Key Takeaways: Non-linear Models

1. **LPM is simple but limited:**
   - Direct probability interpretation
   - Can predict outside [0,1]
   - Use robust SEs
2. **Logit respects [0,1] bounds:**
   - Coefficients are log-odds
   - Must compute marginal effects for interpretation
   - Effects vary by baseline probability
3. **Interactions in logit are complex:**
   - Not just the interaction coefficient
   - Visualize predicted probabilities
   - Effect can change sign across probability range
4. **Report clearly:** State your model choice, show marginal effects or predicted probabilities, not just raw coefficients

Day 3 – Outline

1. Understanding Interaction Effects
2. Non-linear Models and Their Interactions
3. Analysing Nested Data: Country Effects
4. Survey Weighting

Nested Survey Data

The Problem

Observations are **not independent** when nested within countries (or regions, schools, etc.)

Our Example:

- Respondents nested within countries: Great Britain (GB), Germany (DE), France (FR)
- Outcome: Regular news consumption
- Individual predictors: age, gender, education

Key Question: How do we account for unobserved country-level differences?

Country Fixed Effects (Dummies)

Strategy

Add country dummy variables to absorb time-invariant, unobserved differences

Model specification:

```
news_regular ~ agea + gender + eduyrs + country
```

What this does:

- Creates dummy variables for each country (reference = DE, alphabetically first)
- Estimates average gaps relative to the reference country
- Controls for all unobserved, country-specific factors

Fixed Effects: Interpretation

Country coefficients capture:

- Average difference in log-odds of regular news consumption
- Relative to the reference country (Germany)
- *After* controlling for age, gender, and education

Example

If $\beta_{\text{GB}} = 0.35$ (p ¡ 0.05):

- British respondents have 0.35 higher log-odds of regular news consumption than Germans
- Holding age, gender, and education constant

Limitation: Each country estimate is independent—no information sharing across countries

Multilevel Logistic Regression

Strategy

Random intercepts by country with **partial pooling**

Model specification:

```
news_regular ~ agea + gender + eduyrs + (1 | country)
```

What `(1 | country)` means:

- Each country gets its own intercept
- Intercepts are drawn from a common normal distribution
- Extreme country estimates are **shrunk** toward the grand mean

Why Partial Pooling?

Advantages over fixed effects:

1. **Reduces noise:** Small samples benefit from borrowing strength across countries
2. **Improves stability:** Out-of-sample predictions are more reliable
3. **Variance decomposition:** Quantifies how much variation is between vs. within countries
4. **Generalizability:** Treats countries as a random sample from a larger population

Key Insight

Multilevel models recognize that countries differ, but assume those differences come from a common distribution—not treating each country as completely unique

Mathematical Intuition

Fixed Effects:

$$\text{logit}(p_{ij}) = \beta_0 + \sum_k \beta_k X_{ijk} + \gamma_j \cdot \mathbb{I}(\text{country}_j)$$

Each γ_j is estimated independently

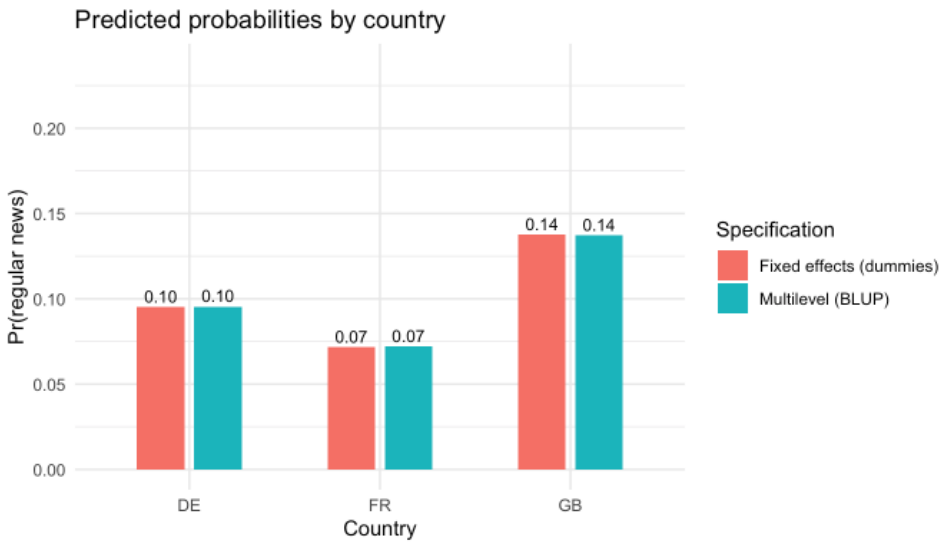
Multilevel Model:

$$\text{logit}(p_{ij}) = \beta_0 + \sum_k \beta_k X_{ijk} + u_j$$

$$u_j \sim N(0, \sigma_u^2)$$

Country effects u_j are drawn from a common distribution with variance σ_u^2

Country Predictions: Fixed vs Multilevel



Typical pattern:

- **Fixed effects:** Wider spread of country estimates
- **Multilevel:** Country estimates closer together (shrunk toward mean)
- Shrinkage is stronger for countries with smaller sample sizes

Interpretation: Multilevel estimates are more conservative because extreme country effects are partially attributed to sampling variation rather than true differences

When to Use Each Approach

Fixed Effects

+ Exact country differences
+ No distributional assumptions
+ Inference about *these* countries only
– Inefficient with many groups
– Unstable with small samples

Multilevel Model

+ Efficient estimation
+ Stable predictions
+ Variance decomposition
+ Generalizes beyond sample
– Assumes normality of effects

Rule of Thumb

With few groups (3–5): Both work, fixed effects slightly more conservative.
With many groups (10+): Multilevel models provide substantial gains.

Random Slopes: Gender by Country

Extension: Allow the gender gap to vary by country

```
ml_logit_gender <- glmer(
  news_regular ~ agea + gender + eduyrs +
                 (1 + gender | country),
  data = ess,
  family = binomial(),
  control = glmerControl(optimizer = "bobyqa")
)
```

What `(1 + gender | country)` adds:

- Random intercepts (baseline by country)
- Random slopes for gender (gender effect varies by country)

Random Slopes: Caution

With Three Countries

Random slopes model may be **over-parameterized**:

- Estimates 2 variance components + 1 correlation = 3 parameters
- Only 3 countries to estimate from
- Risk of convergence issues or unstable estimates

Recommendation for this dataset:

- Stick with random intercepts only: `(1 | country)`
- Consider random slopes when you have 10+ countries
- Always check model convergence and warnings

Key Takeaways

1. **Nested data requires explicit modeling** of clustering (country, region, school, etc.)
2. **Fixed effects**: Control for country differences with dummies
   - Best when: Few groups, focus on these specific countries
3. **Multilevel models**: Partial pooling with random intercepts
   - Best when: Many groups, interest in generalization, variance decomposition
4. **Random slopes**: Allow effects to vary by group
   - Requires sufficient groups (10+) for stable estimation

Practical Recommendations

For your analysis:

- With 3 countries (GB, DE, FR): Either approach is defensible
- Report both if space allows—they tell slightly different stories
- Use random intercepts `(1 | country)` as the default multilevel specification
- Avoid random slopes with only 3 countries

Model comparison:

- Compare AIC/BIC between fixed effects and multilevel
- Check ICCs (Intraclass Correlation) to quantify country-level variation
- Report variance components from multilevel model

Day 3 – Outline

1. Understanding Interaction Effects
2. Non-linear Models and Their Interactions
3. Analysing Nested Data: Country Effects
4. Survey Weighting

Survey Weighting — Why and How

The Problem

Survey data are collected with **unequal selection probabilities**. Inference should reflect the design to avoid biased point estimates and standard errors.

Key Concept:

- Let $w_i = 1/\pi_i$ be the design (inverse-probability) weight
- For a finite population mean: $\bar{Y} = \sum_i w_i Y_i / \sum_i w_i$
- For regression: weighted likelihoods re-scale each case by w_i

ESS fields used:

- `pweight` (post-stratification weight)
- `psu` (primary sampling unit)
- `stratum` (strata)

What These Design Variables Mean

pweight (post-stratification weight):

- Adjusts for unequal inclusion probabilities
- Aligns achieved sample with known population margins (e.g., age $\times$ gender $\times$ region)
- Large values = upweight under-represented respondents
- Small values = downweight over-represented respondents

psu (primary sampling unit):

- First cluster stage of selection (e.g., municipalities, postcode sectors)
- Respondents in same PSU share fieldwork and selection features
- Their responses are **correlated** — affects standard errors

Design Variables (continued)

stratum (strata):

- Mutually exclusive groups within which PSUs were sampled
- Examples: region $\times$ urbanicity
- Stratification improves precision
- Variance estimation must respect it

Design Degrees of Freedom

With clustering and stratification, effective df are closer to the number of PSUs minus strata, **not** the raw respondent count — hence the importance of design-aware SEs.

R Setup:

- Set `options(survey.lonely.psu = "adjust")` to stabilize single-PSU strata

When to Weight: Practical Guidance

USE design/post-strat weights when:

- Estimating population levels (means, totals, prevalence)
- Estimating effects that might shift with differential selection
- Making population-level inferences from sample data

Weights may be OPTIONAL when:

- Working with randomized experiments
- Modeling causal effects with ignorable sampling
- Note: Cluster-robust SEs still matter even without weights

Weights can be SKIPPED when:

- Research question is sample-only prediction
- **Important:** Report that scope of inference is limited to sample

Weighted Regression: Linear Probability Model

Example: Linear probability model with survey weights

```
# Define survey design
des <- svydesign(
  ids = ~psu,
  strata = ~stratum,
  weights = ~pweight,
  data = ess
)

# Weighted linear probability model
lpm_w <- svyglm(
  news_regular ~ agea + gender + country,
  design = des,
  family = gaussian()
)

tidy(lpm_w)
```

Weighted Regression: Logistic Model

Example: Logistic regression with survey weights

```
# Weighted logistic regression
logit_w <- svyglm(
  news_regular ~ agea + gender + country,
  design = des,
  family = quasibinomial()
)

tidy(logit_w)
```

Key Difference

Use `quasibinomial()` family for logistic regression with survey weights (not `binomial()`) to properly account for design effects.

Interpreting Weighted Results

What changes with weighting:

1. **Point estimates:** Coefficients shift to reflect population composition, not just sample composition

2. **Standard errors:** Usually increase due to:
   - Clustering (correlation within PSUs)
   - Unequal weights (design effect)
   - Stratification adjustments

3. **Confidence intervals:** Wider intervals reflect true uncertainty in population inference

4. **P-values:** May change significance of coefficients

Key Takeaways: Survey Weighting

1. **Always use weights** when making population-level inferences from survey data
2. **Specify the full design:** weights + clustering + stratification
   - Missing any component leads to incorrect SEs
3. **Use appropriate functions:**
   - `svydesign()` to define the survey design
   - `svyglm()` for weighted regression
   - `quasibinomial()` for logistic regression
4. **Report your approach:** State whether/how you weighted and justify the decision
5. **Effective df $\neq$ sample size:** With clustering, your effective sample is smaller than the raw count

Practical Considerations

Common mistakes to avoid:

- Using `lm()` or `glm()` with weights argument only (ignores clustering/stratification)
- Forgetting to set `survey.lonely.psu` option
- Using `binomial()` instead of `quasibinomial()` for weighted logit
- Not reporting the weighting scheme in your methods section

Best practices:

- Compare weighted vs. unweighted estimates as sensitivity check
- Report design effects: $deff = \text{Var}_{\text{weighted}}/\text{Var}_{\text{SRS}}$
- Check that weights are reasonable (e.g., no extreme outliers)
- Document all design specifications clearly

Questions?

Next: Hands-on lab with ESS data