UNIVERSITY OF MANNHEIM
School of Social Sciences

Quantitative Methods in Political Science: Linear Regression: Interpreting Substantive Effects via the Simulation Method

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 7 - 20 October 2021

The Course

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Overview: Week 7

Midterm

Simulation-based Inference

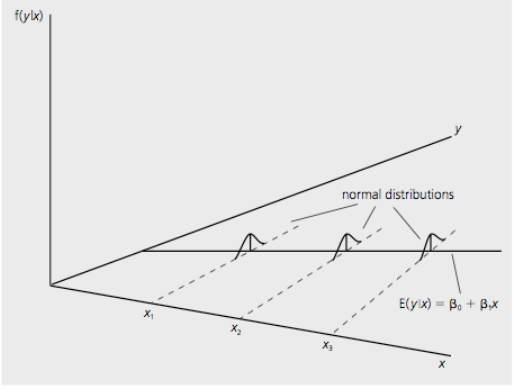
The Simulation Approach

Simulation in Practice

- Example

Quiz

Which of the following statements relating to the assumption of the OLS model (aka Gauss-Markov *assumptions) are correct?*



1. *Across repeated samples X is normally distributed.*
2. *Our model does on average correctly predict the mean of Y given X.*
3. *X is* trichotomous, *i.e. it has only three different values.*
4. *Across repeated samples the errors are identically distributed independent of the values of X.*
5. *The distribution of the errors has the same variance as the distribution of Y conditional on X.*

Midterm

Logistics

- The midterm will take place Wednesday, 3 November 2021, 8:30-10:00, B144.
- We will start 8:30 sharp, so be on time.
- Just bring a pen. There is no need for a notebook; your exam package includes extra pages.
- All of the material covered is relevant.
- Questions will include:
  - Definitions of concepts as required in the homework.
  - Questions on basic statistics.
  - Multiple-choice questions (only one correct answer).
  - Questions about OLS regression.
- If you are required to calculate some quantities, make sure that we can understand where the numbers come from.
- If you are asked to calculate 95% confidence intervals, a rough approximation, i.e., $\hat{\beta} \pm 2 \cdot SE(\hat{\beta})$, is fine.
- We will upload a mock exam to ILIAS.

Simulation-based Inference

Statistical Inference in Linear Models

- Last lecture: Classical statistical regression inference including
  - confidence intervals for estimated coefficients,
  - significance tests for estimated coefficients using confidence interval, t-test, p-values and
- This lecture: Interpretation of regression inference including
  - how to make results accessible to non-technical readers,
  - how to learn about quantities of interest,
  - how to display uncertainty of own results, and
  - which tools to use (predicted values, expected values, and first differences).

Quantities of Interest

- Apart from tests of statistical significance, we usually want to present quantities of interest, which illustrate the meaning and substantive significance of our statistical model.
- Substantive quantities are easier to interpret than raw regression coefficients and avoid technical jargon.
- Compare the two statements:
  - *"The coefficient of income on campaign contribution is 0.25 and statistically significant at the five percent level."*
  - *"If a respondent's income rises by US$ 1,000, we expect her campaign contributions to increase by US$ 250 $\pm$ US$ 50."*

Quantities of Interest

- Presenting quantities of interests (QoI) means to express estimation results in substantive terms (e.g., in units of the dependent variable or as probability of an event). This includes calculating and reporting our uncertainty about these quantities.
- A QoI is a function of the estimated coefficients (and the respective uncertainty).
- Presenting substantive effects is a sign of good empirical practice!
  - It broadens your readership as it allows non-technical readers to understand your results.
  - It helps you to reflect on your findings and to put them into perspective.

Strategy for Substantive Interpretation

1. We get QoI as a function of estimated coefficients
2. Where is the uncertainty in a statistical model?
3. Use simulation to account for estimation and fundamental uncertainty
4. Create plots and tables for communicating your results

Where is the Uncertainty?

Recall that we can write a linear regression model as

$$
\begin{aligned}
Y_i &\sim N(y_i|\mu_i, \sigma^2) && \text{stochastic} \\
\mu_i &= X_i\beta = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots && \text{systematic}
\end{aligned}
$$

1. Estimation Uncertainty: Uncertainty about what the true parameters β and σ^2 of the model are. Think of it as caused by small samples. Vanishes if N gets larger.
2. Fundamental Uncertainty: Represented by stochastic component of the model. Exists no matter what (even if model is correct and we would have infinite many observations and no measurement error) because of inherent randomness of the world.

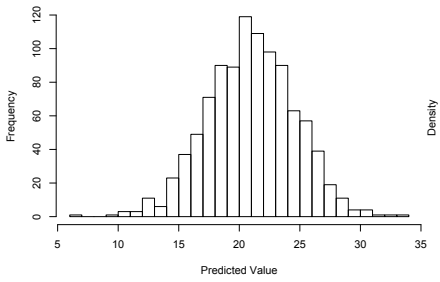
Different Types of Quantities of Interest

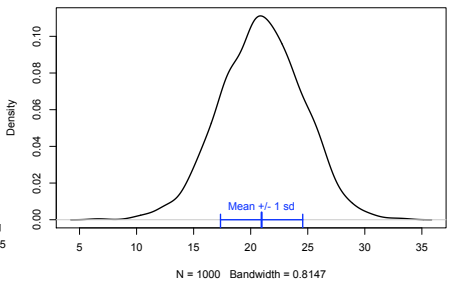
- There are different types of quantities of interest, e.g., ...
  - Marginal effects, $\Delta Y / \Delta X$: How does the DV change if the IV changes and all else is held constant?
  - Predicted values, $\hat{Y} \mid X$: Which value does our model predict, given a particular set of X values?
  - Expected values, $E(Y \mid X)$: Which value of Y do we expect from the model, given a particular set of X values?
  - First differences, $E(Y \mid X_1) - E(Y \mid X_2)$: What is the 'causal' effect (difference in expectations) when we change the set of X values from X_1 to X_2?
  - Anything you want (or your theory would suggest), as long as it is a function of the estimated parameters of the model ...
- The simulation approach has two major advantages:
  - You can simulate any quantity of interest you care about.
  - Since the simulation approach generates distributions, we get uncertainty intervals for free.

The Simulation Approach

Basic Principles: Simulation from Regression Output

1. Set up a sampling distribution for your regression coefficients through simulation.
2. Generate quantity of interest from your model with covariates at specific values (mostly mean for continuous and median for binary variables)
3. Summarize empirical distribution of this newly generated sample to get quantity of interest and uncertainty.

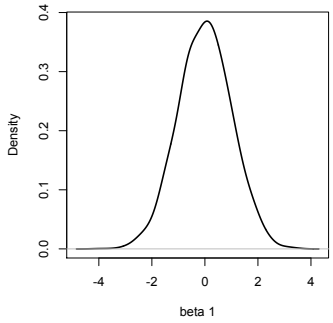




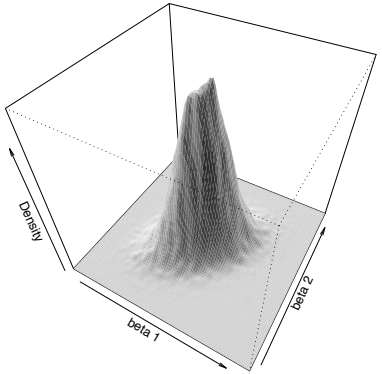
Univariate and Bivariate Normal Distributions

- To model estimation uncertainty we draw from normal distributions with the coefficient as mean and standard error as standard deviation.

$$S \sim N(0, 1)$$



$$S \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$$



Simulation in Practice

Simulating Quantities of Interest: Five Steps

1. Get regression coefficients.
2. Generate sampling distribution to account for estimation uncertainty.
3. Choose covariate values that will be fixed during the simulation.
4. Calculate quantities of interest, such as predicted values, expected values or first differences.
5. Calculate summary measures from simulated distribution of your quantity of interest.

Step 1: Get Regression Coefficients

- Fit a linear model using OLS. This provides us with two central pieces of information:
  - A vector of regression coefficients for the intercept and slope parameters.
  - A variance-covariance matrix that captures uncertainty around coefficients.
- For example, consider our standard linear model for income;

$$Income = \beta_0 + \beta_1 educ + \beta_2 female + \varepsilon,$$

or more compact:

$$y = X\beta + \varepsilon$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24475.06     666.30   36.73   <2e-16 ***
education    1046.84      73.81   14.18   <2e-16 ***
female      -3994.81     294.80  -13.55   <2e-16 ***
```

Step 1: Get Regression Coefficients

- For notational convenience, we can stack all estimated coefficients into a vector, $\hat{\beta}$.
- Similarly, we can write a matrix, $\hat{\mathrm{V}}$, which contains the variances of the coefficients on the main diagonal and the covariances between the coefficients on the off-diagonal. We call such a matrix a variance-covariance matrix.
- We get:

$$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2]', \ \hat{\mathrm{V}} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0\hat{\beta}_1) & Cov(\hat{\beta}_0\hat{\beta}_2) \\ Cov(\hat{\beta}_1\hat{\beta}_0) & Var(\hat{\beta}_1) & Cov(\hat{\beta}_1\hat{\beta}_2) \\ Cov(\hat{\beta}_2\hat{\beta}_0) & Cov(\hat{\beta}_2\hat{\beta}_1) & Var(\hat{\beta}_2) \end{bmatrix}$$

- Hence, $SE(\hat{\beta}) = \sqrt{diag(\hat{\mathrm{V}})}$.
- With our example, $\hat{\beta}$ and $\hat{\mathrm{V}}$ look like this

$$\hat{\beta} = [24475, 1047, -3995]', \ \hat{\mathrm{V}} = \begin{bmatrix} 443949 & -46303 & -55163 \\ -46303 & 5447 & 564 \\ -55163 & 564 & 86908 \end{bmatrix}$$

Step 2: Generate Sampling Distribution

- We are interested in the estimation uncertainty around our coefficient estimates. To do so, we draw a large number of values, n (e.g., 1,000 values).
- We usually do not only have one coefficient in our model, but p $(= k + 1)$ coefficients. Thus, we set up a multivariate normal distribution.
- This distribution has the p x 1 vector of coefficients, $\hat{\beta}$, as its mean and a variance given by the estimated p x p variance-covariance matrix, $\hat{\mathrm{V}}$.
- n draws from such a distribution yield an n x p matrix S which is given as

$$\underset{[n \times p]}{\mathrm{S}} = \mathcal{MVN}(\hat{\beta}, \hat{\mathrm{V}})$$

- For our examples this yields:

$$\mathrm{S} = \begin{bmatrix} 23810 & 1102 & -4000 \\ 24329 & 1111 & -4419 \\ \dots & \dots & \dots \\ 24276 & 1088 & -4117 \end{bmatrix}$$

Step 3: Choose Covariate Values

- Let x be a row vector accounting for an intercept and with specific values for two covariates, $\mathrm{x} = [1, x_1, x_2]$.
- Let's use mean values of education and gender, such that x is given as

$$\mathrm{x} = [1, 8.44, 0.58].$$

- In practice, one chooses either reasonable covariate values that represent the population, or conducts several simulation runs with different covariate values to illustrate the implication of the fitted model.
- To summarize the effect of one or several covariates at once, first differences are extremely helpful!

Step 4: Calculate Quantities of Interest

- For each row of S we calculate our quantity of interest.
- For example, if we are interested in the expected value $\hat{Y} = E(Y \mid X)$ we simply compute

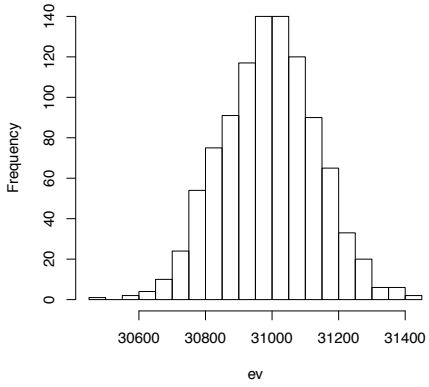
$$\underset{[n\times 1]}{\text{ev}} = \underset{[n\times p]}{\text{S}} \times \underset{[p\times 1]}{\text{x}'}$$

i.e.

$$\begin{bmatrix} 23810 & 1102 & -4000 \\ 24329 & 1111 & -4419 \\ \dots & \dots & \dots \\ 24276 & 1088 & -4119 \end{bmatrix} \times [1, 8.44, 0.58]' = \begin{bmatrix} 30793 \\ 31144 \\ \dots \\ 31074 \end{bmatrix}$$

Step 5: Summarize Results

- The resulting *n x* 1 vector, ev, (e.g., the expected income for a typical respondent) can be plotted and allows us to get means, quantiles and, hence, confidence intervals.



mean(ev) = 30991, sd(ev) = 142

Example: Expected Income for Men and Women

- Expected value, $E(Y|X)$, e.g., expected income for men and women with an average level of education:

Men:	33315
Women:	29307

- Those quantities are easily calculated by hand. But how uncertain are we about these predictions?
- Using the simulation approach we get the following, more informative, table showing expected income with associated 95% confidence intervals (using the 2.5% and the 97.5%-percentile of the simulated sampling distribution):

Men:	33315
	[32860,33761]
Women:	29307
	[28957,29699]

First Differences and Uncertainties I

- Another way to emphasize the difference between men and women, is to ask what size the expected difference in income is:

$$E(income|women) - E(income|men)$$

- Using simulation, we can get an estimate of uncertainty for that difference.
- For this, set two vectors of covariates which differ on your quantity of interest, e.g.:

$$x_{woman} = [1, 8.44, 1] \quad x_{men} = [1, 8.44, 0]$$

- Calculate vector of expected values for men and women:

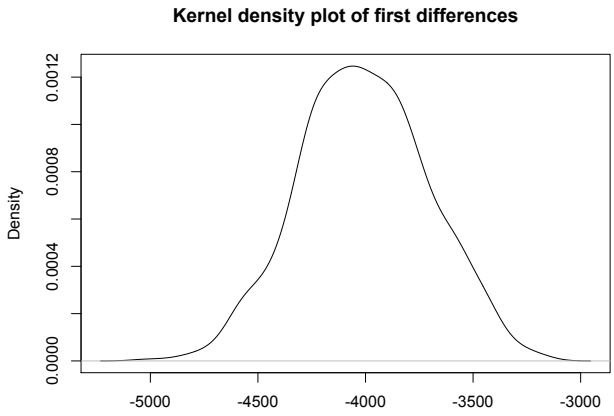
$$ev_{men} = S \times x'_{men} \quad ev_{women} = S \times x'_{women}$$

- The first difference is simply the difference in expected values:

$$\underset{[n\times 1]}{fd} = \underset{[n\times 1]}{ev_{woman}} - \underset{[n\times 1]}{ev_{men}}$$

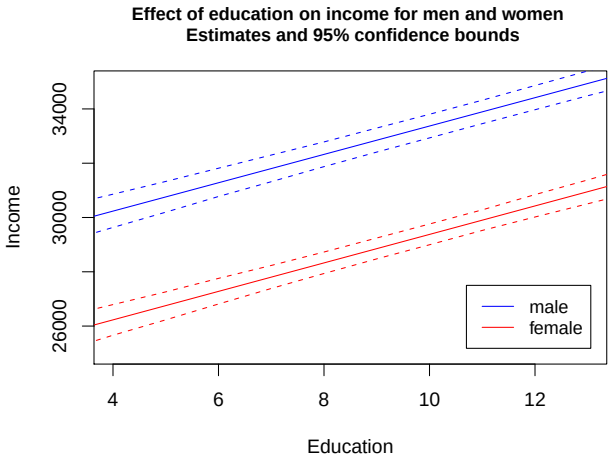
First Differences and Uncertainties II

Difference:	-4007
	[-4588, -3436]



Plot of a Continuous Covariate

- Calculate expected values and confidence bound for each value of a continuous covariate.



How to calculate predicted values?

- Remember, the distribution of predicted values differ from the distribution of expected values only in that it also incorporates fundamental uncertainty.
- To account for that we need to remember what we assumed about the error term. Thus, for every expected value we simply add a random draw from $N(0, \hat{\sigma}^2)$ to account for fundamental uncertainty.
- For example, if we are interested in a concrete predicted value $\hat{Y} \mid X$ we simply compute

$$\underset{[n\times 1]}{\text{pv}} = \underset{[n\times p]}{\text{S}} \times \underset{[p\times 1]}{\text{x}'} + \underset{[n\times 1]}{\text{e}}$$

- The standard errors as well as the confidence intervals for predicted values are larger than for expected values because they also account for fundamental uncertainty of the model.