Interpreting Substantive Effects via the Simulation Method

Lukas F. Stoetzer

Zurich Summer School for Women in Political Methodology

7 August 2017

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	Model				
	(1) Civil War	(2) "Ethnic" War	(3) Civil War	(4) Civil War (Plus Empires)	(5) Civil War (COW)
Prior war	−0.954**	−0.849*	−0.916**	−0.688**	−0.551
	(0.314)	(0.388)	(0.312)	(0.264)	(0.374)
Per capita income[a,b]	−0.344***	−0.379***	−0.318***	−0.305***	−0.309***
	(0.072)	(0.100)	(0.071)	(0.063)	(0.079)
log(population)[a,b]	0.263***	0.389***	0.272***	0.267***	0.223**
	(0.073)	(0.110)	(0.074)	(0.069)	(0.079)
log(% mountainous)	0.219**	0.120	0.199*	0.192*	0.418***
	(0.085)	(0.106)	(0.085)	(0.082)	(0.103)
Noncontiguous state	0.443	0.481	0.426	0.798**	−0.171
	(0.274)	(0.398)	(0.272)	(0.241)	(0.328)
Oil exporter	0.858**	0.809*	0.751**	0.548*	1.269***
	(0.279)	(0.352)	(0.278)	(0.262)	(0.297)
New state	1.709***	1.777***	1.658***	1.523***	1.147**
	(0.339)	(0.415)	(0.342)	(0.332)	(0.413)
Instability[a]	0.618**	0.385	0.513*	0.548*	0.584*
	(0.235)	(0.316)	(0.242)	(0.225)	(0.268)
Democracy[a,c]	0.021	0.013			
	(0.017)	(0.022)			
Ethnic fractionalization	0.166	0.146	0.164	0.490	−0.119
	(0.373)	(0.584)	(0.368)	(0.345)	(0.396)
Religious fractionalization	0.285	1.533*	0.326		1.176*
	(0.509)	(0.724)	(0.506)		(0.563)
Anocracy[a]			0.521*		0.597*
			(0.237)		(0.261)
Democracy[a,d]			0.127		0.219
			(0.304)		(0.354)
Constant	−6.731***	−8.450***	−7.019***	−6.801***	−7.503***
	(0.736)	(1.092)	(0.751)	(0.681)	(0.854)
N	6327	5186	6327	6360	5378

Note: The dependent variable is coded "1" for country years in which a civil war began and "0" in all others. Standard errors are in parentheses. Estimations performed using Stata 7.0. *$p < .05$; **$p < .01$; ***$p < .001$.
[a] Lagged one year.
[b] In 1000's.
[c] Polity IV; varies from −10 to 10.
[d] Dichotomous.

Figure: Regression Results in Fearon and Laitin (2003)

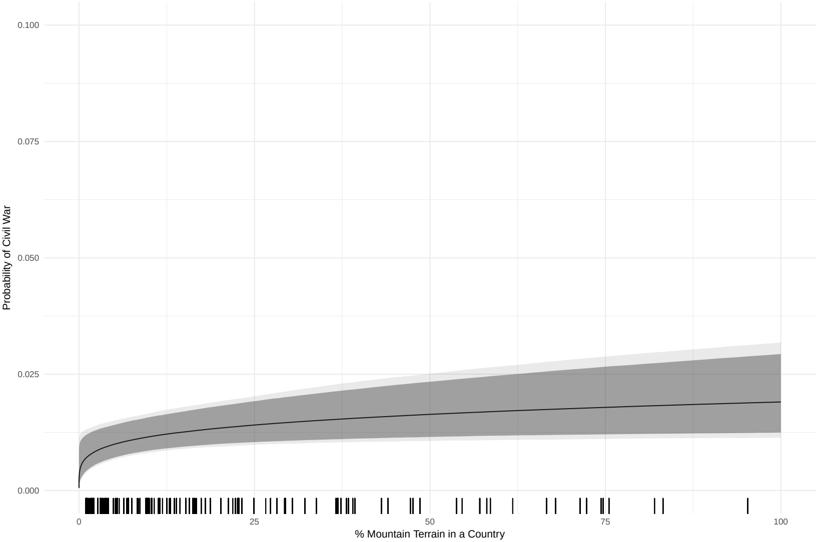


Figure: The effect of Mountain Terrain on the Probability of Civil War

Statistical Inference in Linear Models

- Interpretation of regression inference using simulation
  - 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 probabilities, 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.

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

- 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 ± US$ 50.*"
- Presenting quantities of interests means to express estimation results in substantive terms in units of the dependent variable. This includes calculating and reporting our uncertainty about these quantities.

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 ± US$ 50.*"
- Presenting quantities of interests means to express estimation results in substantive terms in units of the dependent variable. This includes calculating and reporting our uncertainty about these quantities.
- 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

Strategy for Substantive Interpretation

1. We get QoI as a function of estimated coefficients
2. Where is the uncertainty in a statistical model?

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

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

The structure of Generalized Linear Models

A *Generalized Linear Model* (GLM) consist three components

1. *Stochastic Component*, specifying the conditional distribution of the dependent variable Y_i
2. *Systematic Component*, a linear function of predictors, e.g.,

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} = X_i \beta \tag{1}$$

3. *Link Function* $g(\cdot)$ which transforms the expectation of the dependent variable, $\mu_i = E(Y_i)$, to the linear predictor

$$g(\mu_i) = \eta_i = X_i \beta \tag{2}$$

4. The inverse of the *Link Function* $g(\cdot)$ is the *Response Function* $h(\cdot)$.

$$\mu_i = h(\eta_i) = h(X_i \beta) \tag{3}$$

Examples of Generalized Linear Models

- Some of applied models:

Model	Distribution	Link $\eta_i = g(\mu_i)$	Response $\mu_i = h(\eta_i)$	Range Y_i
Linear	Gaussian	μ_i	η_i	$(-\infty, \infty)$
Logit	Bernoulli	$log \frac{\mu_i}{1-\mu_i}$	$\frac{exp(\eta_i)}{1+exp(\eta_i)}$	$[0, 1]$
Probit	Bernoulli	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$	$[0, 1]$
Multinomial	Multinomial	$log \frac{\mu_{ik}}{\sum_{k=1}^{K} \mu_{ik}}$	$\frac{exp(\eta)}{\sum_{k=1}^{K} exp(\eta_{ik})}$	$[0, 1]$
Poisson	Poisson	$log(\mu_i)$	$exp(\eta_i)$	$\mathbb{N}_+$

Where is the Uncertainty?

Generalized Linear models can be written:

$$
\begin{aligned}
Y_i &\sim p(y_i|\mu_i, \alpha) && \text{stochastic} \\
g(\mu_i) &= X_i\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots && \text{systematic}
\end{aligned}
$$

Where is the Uncertainty?

Generalized Linear models can be written:

$$
\begin{aligned}
Y_i &\sim p(y_i|\mu_i, \alpha) && \text{stochastic} \\
g(\mu_i) &= X_i\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots && \text{systematic}
\end{aligned}
$$

1. Estimation Uncertainty: Uncertainty about what the true parameters β and α of the model are. Think of it as caused by small samples. Vanishes if N gets larger.

Where is the Uncertainty?

Generalized Linear models can be written:

$$
\begin{aligned}
Y_i \quad &\sim \quad p(y_i|\mu_i, \alpha) && \text{stochastic} \\
g(\mu_i) \quad &= \quad X_i\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots && \text{systematic}
\end{aligned}
$$

1. Estimation Uncertainty: Uncertainty about what the true parameters β and α 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 measuement 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 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 ...

Different Types of Quantities of Interest

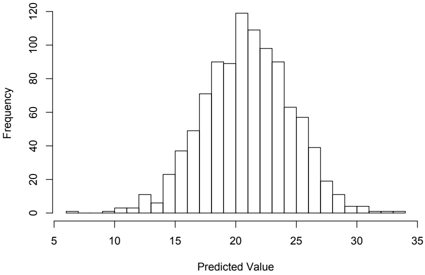
- There are different types of quantities of interest,e.g., ...
  - Marginal effects, $\Delta Y / \Delta X$: How does the DV change if the IV changes and all else is held constant?
  - Predicted values, $\hat{Y} \mid X$: Which value does our model predict, given a particular set of X values?
  - Expected values, $E(Y \mid X)$: Which value 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 ...

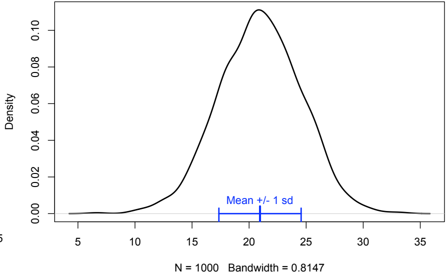
Different Types of Quantities of Interest

- 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 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 that you want to.
  - Since the simulation approach generates distributions, we get uncertainty bounds for free.

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.

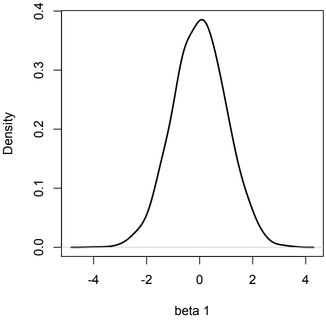




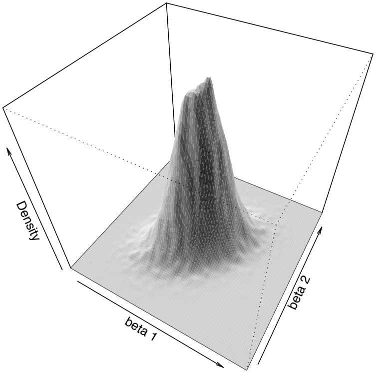
Univariate and Bivariate Normal Distributions

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

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



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



The Magic Five Steps of Simulation

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

Step 1: Get Regression Coefficients

- Fit a regression model. This provides us with two central pieces of information:
  - A vector of regression coefficients.
  - A variance-covariance matrix that captures uncertainty around coefficients.

Step 1: Get Regression Coefficients

- Fit a regression model. This provides us with two central pieces of information:
  - A vector of regression coefficients.
  - 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:

$$\mathrm{y} = \mathrm{X}\beta + \varepsilon$$

```
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, β.
- Similarly, we can write a matrix, 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.

Step 1: Get Regression Coefficients

- For notational convenience, we can stack all estimated coefficients into a vector, β.
- Similarly, we can write a matrix, 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:

$$\beta = [\beta_0, \beta_1, \beta_2]', \ \mathrm{V} = \begin{bmatrix} Var(\beta_0) & Cov(\beta_0\beta_1) & Cov(\beta_0\beta_2) \\ Cov(\beta_1\beta_0) & Var(\beta_1) & Cov(\beta_1\beta_2) \\ Cov(\beta_2\beta_0) & Cov(\beta_2\beta_1) & Var(\beta_2) \end{bmatrix}$$

Step 1: Get Regression Coefficients

- For notational convenience, we can stack all estimated coefficients into a vector, β.
- Similarly, we can write a matrix, 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:

$$\beta = [\beta_0, \beta_1, \beta_2]', \ \mathrm{V} = \begin{bmatrix} Var(\beta_0) & Cov(\beta_0\beta_1) & Cov(\beta_0\beta_2) \\ Cov(\beta_1\beta_0) & Var(\beta_1) & Cov(\beta_1\beta_2) \\ Cov(\beta_2\beta_0) & Cov(\beta_2\beta_1) & Var(\beta_2) \end{bmatrix}$$

- Hence, $SE(\beta) = \sqrt{diag(\mathrm{V})}$.

Step 1: Get Regression Coefficients

- For notational convenience, we can stack all estimated coefficients into a vector, β.
- Similarly, we can write a matrix, 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:

$$\beta = [\beta_0, \beta_1, \beta_2]', \ \mathrm{V} = \begin{bmatrix} Var(\beta_0) & Cov(\beta_0\beta_1) & Cov(\beta_0\beta_2) \\ Cov(\beta_1\beta_0) & Var(\beta_1) & Cov(\beta_1\beta_2) \\ Cov(\beta_2\beta_0) & Cov(\beta_2\beta_1) & Var(\beta_2) \end{bmatrix}$$

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

$$\beta = [24475, 1047, -3995]', \ \mathrm{V} = \begin{bmatrix} 443949 & -46303 & -55163 \\ -46303 & 85447 & 564 \\ -55163 & 564 & 86908 \end{bmatrix}$$

Step 2: Choose Covariate Values

- Let x be a row vector accounting for an intercept and with specific values for two covariates, $\mathrm{x} = [1, x_1, x_2]$.

Step 2: 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

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

Step 2: 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 3: Generate Sampling Distribution

- We are interested in the uncertainty around our coefficient estimates. To do so, we draw a large number of values, n (e.g., 1,000 values).

Step 3: Generate Sampling Distribution

- We are interested in the 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.

Step 3: Generate Sampling Distribution

- We are interested in the 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, β, as its mean and a variance given by the estimated p x p variance-covariance matrix, V.

Step 3: Generate Sampling Distribution

- We are interested in the 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, β, as its mean and a variance given by the estimated p x p variance-covariance matrix, V.
- n draws from such a distribution yield an n x p matrix S which is given as

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

Step 3: Generate Sampling Distribution

- We are interested in the 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, β, as its mean and a variance given by the estimated p x p variance-covariance matrix, V.
- n draws from such a distribution yield an n x p matrix S which is given as

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

- For our examples this yields:

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

Step 4: Calculate Quantities of Interest

- For each row of S we calculate our quantity of interest.

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

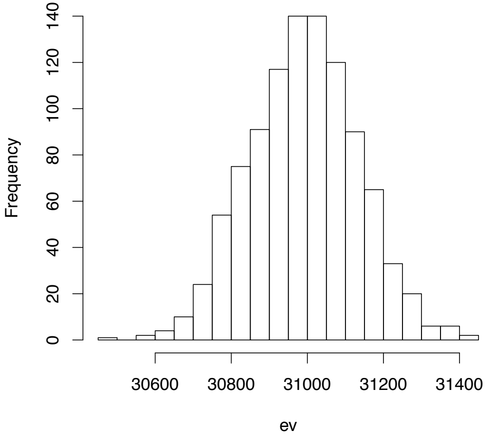
$$\underset{[n \times 1]}{\text{ev}} = \underset{[n \times p]}{\text{S}} \times \underset{[p \times 1]}{\text{x}'}$$

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 $n \times 1$ vector, ev, can be plotted and allows us to get means and quantiles etc.



$$\text{mean(ev)} = 30991, \text{ sd(ev)} = 142$$

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

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?

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 bounds (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:

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

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:

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

- 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.:

$$\mathrm{x}_{men} = [1, 8.44, 0] \quad \mathrm{x}_{women} = [1, 8.44, 1]$$

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:

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

- 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.:

$$\mathrm{x}_{men} = [1, 8.44, 0] \quad \mathrm{x}_{women} = [1, 8.44, 1]$$

- Calculate vector of expected values for men and women:

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

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:

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

- 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.:

$$\mathrm{x}_{men} = [1, 8.44, 0] \quad \mathrm{x}_{women} = [1, 8.44, 1]$$

- Calculate vector of expected values for men and women:

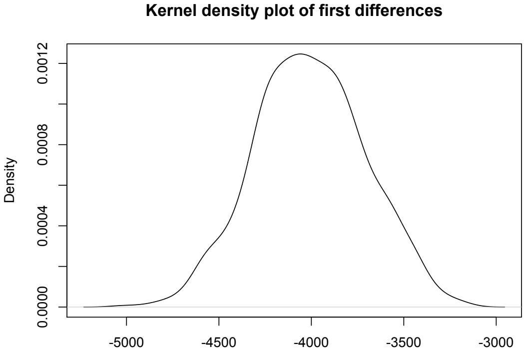
$$\mathrm{ev}_{men} = \mathrm{S} \times \mathrm{x}'_{men} \quad \mathrm{ev}_{women} = \mathrm{S} \times \mathrm{x}'_{women}$$

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

$$\underset{[n \times 1]}{\mathrm{fd}} = \underset{[n \times 1]}{\mathrm{ev}_{men}} - \underset{[n \times 1]}{\mathrm{ev}_{women}}$$

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.