Partial Effects

We want to explain some variable y, called the **explained variable**, using the **explanatory variables** $\mathbf{x}$. We will often consider the effect that explanatory variable w has on the conditional expectation of y, holding fixed a vector of **controls** denoted $\mathbf{c}$. In other words, we are interested in $E[y|w, \mathbf{c}]$.

More generally, we have will have $E[y|\mathbf{x}]$, where $\mathbf{x}$ includes the variable of interest as well as the controls. We will use a **parametric model**. For instance, we might write

$$E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{1}$$

The **partial effect** x_j on $E[y|\mathbf{x}]$ is the partial derivative of $E[y|\mathbf{x}]$ with respect to x_j, and the change in $E[y|\mathbf{x}]$ can be written as

$$\Delta E[y|\mathbf{x}] \approx \frac{E[y|\mathbf{x}]}{\partial x_j} \Delta x_j.$$

For instance, looking at equation (1), we have

$$\frac{E[y|\mathbf{x}]}{\partial x_1} = \beta_1.$$

Thus, if we compare data with a difference of one unit of x_1, then we'd expect the data with the higher unit of x_1 to be larger by β_1. Depending on how the conditional expectation is specified, this partial effect need not be a constant – it could depend on any number of combinations of the explanatory variables.

The most common way to define the **elasticity** of $E[y|\mathbf{x}]$ with respect to x_j is

$$\xi_j = \frac{\partial E[y|\mathbf{x}]}{E[y|\mathbf{x}]} \frac{x_j}{\partial x_j},$$

which gives the percentage change in $E[y|\mathbf{x}]$ per differential percentage change in x_j. This could equivalently be written as

$$\xi_j = \frac{\partial \log\big(E[y|\mathbf{x}]\big)}{\partial \log(x_j)}.$$

There are some contexts where perhaps one would want to define elasticity to be

$$\xi_j^* = \frac{\partial E[\log(y)|\mathbf{x}]}{\partial \log(x_j)}.$$

These are not equivalent in general, but are when u is independent of $\mathbf{x}$. I am not a fan of the latter definition because it is new to me and I fear new things.

Sometimes we might want to consider the percentage change in $E[y|\mathbf{x}]$ from one **unit** change in x_j. This is the **semielasticity** of $E[y|\mathbf{x}]$ with respect to x_j, calculated as

$$100 \times \frac{\partial \log\big(E[y|\mathbf{x}]\big)}{\partial x_j}.$$

Intuition: logs mean percentages, absence of logs means units.

When we consider the expectation of y given data $\mathbf{x}$, it is not always true that we will get the true value of y – we are just looking at an average. Thus, there will usually be some error. Thus, we have

$$y = E[y|\mathbf{x}] + u,$$

where u is the **error term**. Furthermore, we will need $E[u|\mathbf{x}] = 0$ for our parametric model. Consider the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$$

In order to get equation (1) from this, we have

$$E[y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + E[u|\mathbf{x}].$$

So $E[u|\mathbf{x}]$ was a function of x_1 or x_2, then the partial effects would be wrong; if it was a constant, then the intercept term β_0 would be wrong.

Oaxaca-Blinder Decomposition

$$\begin{aligned} E[Y|s=1] - E[Y|s=0] = E[\mu_1(x)|s=1] - E[\mu_0(x)|s=1] \\ + E[\mu_0(x)|s=1] - E[\mu_0(x)|s=0] \end{aligned}$$

Independence

$u \perp \mathbf{x}$ means that $\mathbf{x}$ and u are independent.

(a) If $u \perp \mathbf{x}$, then $E[u|\mathbf{x}] = E[u]$.

(b) But $E[u|\mathbf{x}] = E[u]$ does not imply that $u \perp \mathbf{x}$.

(c) And $E[u|\mathbf{x}] = 0$ does not imply $u \perp \mathbf{x}$.

(d) If $E[u] = 0$ and $u \perp \mathbf{x}$, then $E[u|\mathbf{x}] = 0$.

(e) If $E[u|\mathbf{x}] = 0$, then $E[uf(\mathbf{x})] = 0$.

(f) $E[uf(\mathbf{x})] = 0$ does not imply that $E[u|\mathbf{x}] = 0$.

Law of Iterated Expectation

From $E[u|\mathbf{x}] = 0$, the **law of iterated expectation** implies that $E[u] = 0$ because

$$E[u] = E\big[E[u|\mathbf{x}]\big] = E[0] = 0.$$

Furthermore,

$$Cov(x_j, u) = E[x_j u] - E[x_j]E[u] = E[x_j u].$$

Again from the law of iterated expectation, we have

$$E[x_j u] = E\big[E[x_j u|x]\big] = E\big[x_j E[u|x]\big] = 0.$$

Thus, we have $E[u] = 0$ and $Cov(x_j, u) = 0$.

We can go further and state that u is uncorrelated with *any* function of $\mathbf{x}$ because $E[g(x)]E[u] = 0$ and

$$E[g(\mathbf{x})u] = E\big[E[g(\mathbf{x})u|\mathbf{x}]\big] = E\big[g(\mathbf{x})E[u|\mathbf{x}]\big] = 0.$$

This is useful because it means $E[y|x]$ is properly specified. Suppose we'd assumed only $E[xu] = 0$. This means there is no more *linear* information in $\mathbf{x}$ that helps predict y. But if it wasn't also the case that $E[g(\mathbf{x})u] = 0$, then there would be some nonlinear information in $\mathbf{x}$ that would help us to predict y that we'd omitted.

Suppose $\mathbf{x} = \mathbf{f}(\mathbf{w})$. A more general way of writing the law of iterated expectations is

$$E[y|\mathbf{x}] = E\big[E[y|\mathbf{w}]\big|\mathbf{x}\big],$$

as well as

$$E[y|\mathbf{x}] = E\big[E[y|\mathbf{x}]\big|\mathbf{w}\big].$$

The idea is that $\mathbf{x}$ has no more information, and possibly less information, than $\mathbf{w}$ because $\mathbf{f}(\mathbf{w})$ need not be bijective. Thus, the placement of $\mathbf{x}$ and $\mathbf{w}$ on the RHS doesn't matter.

Sometimes we need the special case (e.g. for the Oaxaca-Blinder decomposition) where

$$E[y|\mathbf{x}] = E\big[E[y|\mathbf{x}, \mathbf{z}]\big|\mathbf{x}\big].$$

Asymptotics

Definition 1. *A sequence $\{a_n\}$ is $\boldsymbol{O(N^\lambda)}$ if $N^{-\lambda}a_N$ is bounded. When $\lambda = 0$ and $\{a_N\}$ is bounded, we write $a_N = O(1)$, pronounced* "big oh one."

Consider $a_N = 10 + \sqrt{N}$. By itself, $a_N \to \infty$ as $N \to \infty$ because $\sqrt{N}$ is unbounded. But if we divide a_N by $N^{1/2}$, then the sequence converges to 1. Thus, a_N is $O(N^{1/2})$. It's also $O(N^\lambda)$ for any $\lambda \geq 1/2$, but usually we just choose the smallest such λ.

Definition 2. *The sequence $\{a_N\}$ is $\boldsymbol{o(N^\lambda)}$ if $N^{-\lambda}a_N \to 0$. When $\lambda = 0$, a_N converges to zero, and we write $a_N = o(1)$, pronounced* "little oh one."

We know that $a_N = 10 + \sqrt{N}$ converges to 1 if $\lambda = 1/2$. If $\lambda > 1/2$, then a_N converges to 0. Thus, a_N is $o(N^{1/2+\epsilon})$ for any $\epsilon > 0$.

Notice that if a_N is $o(N^\lambda)$, then a_N is $O(N^\lambda)$; if the sequence converges to zero, clearly it is bounded.

We write $X_N \xrightarrow{p} a$ when X_N converges in probability to a. When $a = 0$, we say that X_N is $o_p(1)$ and write $X_N = o_p(1)$.

We can also have a sequence of random variables that is bounded in probability, meaning that

$$P(|X_N| \geq b_\epsilon) < \epsilon$$

for large enough N. In this case, we write $X_N = O_p(1)$.

Notice that if $X_N \xrightarrow{p} a$, then $X_N = O_p(1)$.

Definition 3. *A random sequence X_N is $o_P(a_N)$, where $\{a_N\}$ is a nonrandom, positive sequence, if $x_N/a_N = o_p(1)$. We write $X_N = o_p(a_N)$.*

In other words, if we can divide every term in the sequence X_N by the corresponding terms in a_N, and it ends up converging in probability to zero, then it is $o_p(a_N)$.

Definition 4. *A random sequence X_N is $O_p(a_N)$, where $\{a_N\}$ is a nonrandom, positive sequence, if $x_N/a_N = O_p(1)$. We write $x_N = O_p(a_N)$.*

In other words, if we divide every term in X_N by the corresponding term in a_N and the sequence ends up bounded in probability, then X_N is in $O_p(a_N)$.

Abusing notation a little, we have the following results.

(a) $o_p(1) + o_p(1) = o_p(1)$
(b) $O_p(1) + O_p(1) = O_p(1)$
(c) $O_p(1) \times O_p(1) = O_p(1)$
(d) $o_p(1) \times O_p(1) = o_p(1)$

Theorem 1 (Slutsky's Theorem)**.** *Let $\mathbf{g} : \mathbb{R}^K \to \mathbb{R}^J$ be a function that is continuous at some point $\mathbf{c} \in \mathbb{R}^K$. Let $\{\mathbf{x}_N\}$ be a sequence of $K \times 1$ random vectors such that $\mathbf{x}_n \xrightarrow{p} \mathbf{c}$. Then $\mathbf{g}(\mathbf{x}_N) \xrightarrow{p} \mathbf{g}(\mathbf{c})$ as $N \to \infty$.*

In other words, $\operatorname{plim} \mathbf{g}(\mathbf{x}_N) = \mathbf{g}(\operatorname{plim} \mathbf{x}_N)$ *if $\mathbf{g}(\cdot)$ is continuous at* $\operatorname{plim} \mathbf{x}_N$.

I will emphasize the fact that Slutsky's theorem only requires continuity at the point of interest $\mathbf{c}$.

Definition 5. *A sequence of random variables x_N **converges in distribution** to the continuous random variable x if and only if*

$$F_N(\xi) \to F(\xi) \quad \text{as } N \to \infty \text{ for all } \xi \in \mathbb{R},$$

where F_N is the cdf of x_N and F is the continuous cdf of x. We write $x_N \xrightarrow{d} x$.

Notice that no x_N is required to be continuous itself. When $x \sim \mathcal{N}(\mu, \sigma^2)$, we write $x_N \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$ or $x_N \overset{a}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Definition 6. *Suppose $\{\mathbf{x}_N\}$ is a sequence of $K \times 1$ vectors. Then $\{\mathbf{x}_N\}$ converges in distribution to $\mathbf{x}$ if and only if for any $K \times 1$ nonrandom vector $\mathbf{c}$ such that $\mathbf{c}'\mathbf{c} = 1$, then*

$$\mathbf{c}'\mathbf{x}_N \xrightarrow{d} \mathbf{c}'\mathbf{x}.$$

When $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$, we require that $\mathbf{c}'\mathbf{x}_N \xrightarrow{d} \mathcal{N}(\mathbf{c}'\mathbf{m}, \mathbf{c}'\mathbf{V}\mathbf{c})$ for every $\mathbf{c} \in \mathbb{R}^K$. In this case, we write $\mathbf{x}_N \xrightarrow{d} \mathcal{N}(\mathbf{m}, \mathbf{V})$ or $\mathbf{x}_N \overset{a}{\sim} \mathcal{N}(\mathbf{m}, \mathbf{V})$.

If $\mathbf{x}_N \xrightarrow{d} \mathbf{x}$, then $\mathbf{x}_N = O_p(1)$.

Theorem 2 (Continuous Mapping Theorem)**.** *Let $\{\mathbf{x}_n\}$ be a $K \times 1$ sequence of random vectors such that $\mathbf{x}_N \xrightarrow{d} \mathbf{x}$. If $\mathbf{g} : \mathbb{R}^K \to \mathbb{R}^J$ is a continuous function, then $\mathbf{g}(\mathbf{x}_N) = \mathbf{g}(\mathbf{x})$.*

So for convergence in probability via Slutsky's Theorem, we need continuity at only the point of convergence. Whereas for convergence in distribution via the Continuous Mapping Theorem, we need the function to be continuous everywhere.

Theorem 3 (Lindeberg-Levy CLT)**.** *Let $\{\mathbf{w}_i\}$ be i.i.d $J \times 1$ vectors such that $E[w_{ij}^2] < \infty$ for $j = 1, \dots J$ and $E[\mathbf{w}_i] = 0$. Then $\{\mathbf{w}_i\}$ satisfies the central limit theorem,*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_i\right) \xrightarrow{d} \mathcal{N}(0, B),$$

where $B = E[\mathbf{w}_i\mathbf{w}_i']$ is positive semi-definite.

Definition 7. *Let $\{\hat{\boldsymbol{\theta}}_N\}$ be a sequence of estimators of the $P \times 1$ vector $\boldsymbol{\theta}$. Suppose that*

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

*where $\mathbf{V}$ is a $P \times P$ positive semidefinite matrix. Then we say $\hat{\boldsymbol{\theta}}_N$ is **$\sqrt{N}$-asymptotically normally distributed** and $\mathbf{V}$ is the **asymptotic variance** of $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta})$, denoted*

$$\operatorname{Avar}\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) = \mathbf{V}.$$

Note that $\mathbf{V}/N = \operatorname{Var}(\hat{\boldsymbol{\theta}}_N)$ only in special cases, and $\hat{\boldsymbol{\theta}}_N$ is rarely normally distributed. But if we can conclude $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$, then we treat $\hat{\boldsymbol{\theta}}_N$ as if

$$\hat{\boldsymbol{\theta}}_N \sim \mathcal{N}\left(\boldsymbol{\theta}, \frac{\mathbf{V}}{N}\right),$$

and therefore $\operatorname{Avar}(\hat{\boldsymbol{\theta}}) = \mathbf{V}/N$.

Definition 8. *If $\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}) \overset{a}{\sim} \mathcal{N}(0, \mathbf{V})$ where $\mathbf{V}$ is positive semidefinite with jth diagonal v_{jj} and $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{V}$, then the **asymptotic standard error** of θ_{Nj}, denoted $se(\hat{\theta}_{Nj})$, is $\sqrt{\hat{v}_{Njj}/N}$.*

Testing

Suppose we have a null hypothesis $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$, where r is a $Q \times 1$ nonrandom vector. If we want to test the null against $H_1 : \mathbf{R}\boldsymbol{\theta} \neq \mathbf{r}$, we define the **Wald statistic** to be

$$W_N := (\mathbf{R}\hat{\boldsymbol{\theta}}_\mathbf{N} - \mathbf{r})'[\mathbf{R}(\hat{\mathbf{V}}_N/N)\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}}_\mathbf{N} - \mathbf{r}).$$

Under H_0, it turns out that $W_N \sim \mathcal{X}_Q^2$.

Theorem 4 (Delta Method)**.** *Suppose θ is $\sqrt{N}-$asymptotically normal. Further suppose $\mathbf{V}$ is positive definite. Let $\mathbf{c} : \Theta \to \mathbb{R}^Q$ be a continuously differentiable function on the parameter space $\theta \in \mathbb{R}^P$ where $Q \leq P$. Also assume that $\boldsymbol{\theta}$ is in the interior of the parameter space. Define $\mathbf{C}(\boldsymbol{\theta}) = \nabla_\theta \mathbf{c}(\boldsymbol{\theta})$ as the $Q \times P$ Jacobian of $\mathbf{c}$. Then*

$$\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})] \overset{a}{\sim} \mathcal{N}\left(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})'\right).$$

It follows that

$$\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})]'[\mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})']^{-1}\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})]' \overset{a}{\sim} \mathcal{X}_Q^2.$$

Define $\hat{\mathbf{C}} = \mathbf{C}(\boldsymbol{\theta}_N)$. Then $\operatorname{plim}\hat{\mathbf{C}} = \mathbf{C}(\boldsymbol{\theta})$. If $\operatorname{plim}\hat{\mathbf{V}} = \mathbf{V}$, then

$$\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})]'[\hat{\mathbf{C}}_N\hat{\mathbf{V}}_N\hat{\mathbf{C}}_N']^{-1}\sqrt{N}[\mathbf{c}(\hat{\boldsymbol{\theta}}_N) - \mathbf{c}(\boldsymbol{\theta})]' \overset{a}{\sim} \mathcal{X}_Q^2.$$