LEAST SQUARES, ADAPTIVE PARTIALLING-OUT, SIMULTANEOUS INFERENCE

ABSTRACT. Here we overview the least squares from several interesting angles. We discuss Frisch-Waugh-Lovell partialling out and point out its adaptivity property in establishing approximate normality of the regression estimators of a set of target regression coefficients. We then discuss construction of simultaneous confidences sets for this set. We make use of the methods to analyze the gender wage gap and the impact of reemployment incentives on the duration of unemployment.

1. NOTATION

For two sequences of real numbers, $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, the notation $a_n \lesssim b_n$ means there exists C such that for all n we have that $a_n \leq Cb_n$, for some constant C that does not depend on n. For a vector $v = (v_1, v_2, ..., v_k)' \in \mathbb{R}^k$, the ℓ_2 and ℓ_1 norms are denoted by $\|\cdot\|_2$ (or simply $\|\cdot\|$) and $\|\cdot\|_1$, respectively,

$$\|v\|_2 := \left(\sum_{i=1}^{k} v_i^2\right)^{1/2}, \quad \|v\|_1 := \sum_{i=1}^{k} |v_i|.$$

The ℓ_0-"norm", $\|\cdot\|_0$, denotes the number of non-zero components of a vector, and the $\|\cdot\|_\infty$ denotes the max norm:

$$\|v\|_0 := \sum_{i=1}^{k} 1\{v_i \neq 0\}, \quad \|v\|_\infty := \max\{|v_i| : i \in \{1, ..., k\}\}.$$

When applied to a matrix, $\|\cdot\|$ denotes the operator norm, namely

$$\|A\| := \max\{\|Av\| : \|v\| \leq 1\}.$$

We use the notation $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We use x' to denote the transpose of a column vector x. In what follows we use the notion $\mathbb{E}_n f(W)$ abbreviates the empirical expectation of $f(W)$ as W ranges over the sample $(W_i)_{i=1}^n$:

$$\mathbb{E}_n f(W) = \frac{1}{n} \sum_{i=1}^{n} f(W_i),$$

2. Least Squares

Let Y be a scalar random variable and X be a p-vector of covariates called regressors. We observe n i.i.d. copies $\{(Y_i, X_i')\}_{i=1}^n$ of (Y, X'). Note that independence is not needed in many places, as is clear from the context. Throughout we assume that $\mathrm{E}Y^2$ and $\mathrm{E}XX'$ are finite.

We then define least squares or projection parameter β in the *population* as the solution of the following prediction problem:

$$\beta := \arg\min_{b \in \mathbb{R}^p} \mathrm{E}(Y - X'b)^2$$

where β obeys the first-order condition:

$$\mathrm{E}(Y - X'\beta)X = 0,$$

and provided that $\mathrm{E}XX'$ is of full rank, which amounts to absence of the multicollinearity, has the closed form expression:

$$\beta = (\mathrm{E}XX')^{-1}\mathrm{E}XY,$$

Defining $\varepsilon = Y - X'\beta$, we obtain the decomposition identity

$$Y \equiv X'\beta + \varepsilon, \quad \mathrm{E}\varepsilon X = 0.$$

Observe that we did not need any linearity assumption to obtain this decomposition.

We define least squares estimator or projection estimator $\hat{\beta}$ in the *sample* as the solution of the following prediction problem:

$$\hat{\beta} := \arg\min_{b \in \mathbb{R}^p} \mathbb{E}_n(Y - X'b)^2$$

which obeys the first-order condition:

$$\mathbb{E}_n(Y - X'\hat{\beta})X = 0,$$

and has the closed form solution

$$\hat{\beta} = (\mathbb{E}_n XX')^{-1}\mathbb{E}_n XY,$$

provided that $\mathbb{E}_n XX'$ is of full rank, which amounts to absence of the multicollinearity in the *sample*. Defining $\hat{\varepsilon}_i = Y_i - X_i'\hat{\beta}$, we obtain the decomposition identity

$$Y_i \equiv X_i'\hat{\beta} + \hat{\varepsilon}_i, \quad \mathbb{E}_n\hat{\varepsilon}X = 0.$$

Note that the least squares estimator makes sense only if p is not bigger than n. If $p > n$ other estimators must be used, for example, penalized least squares estimators or post-selection least squares estimators.

3. Partialling Out. Frisch-Waugh-Lovell Theorem

This is an important tool that provides conceptual understanding of least squares as well as a very practical tool for estimation and visualization of results. We partition vector of regressors X into two groups:

$$X = (D', W')',$$

where p_1-dimensional subvector D represents "target" regressors of interest, and p_2-dimensional subvector W represents other regressors, sometimes called the controls. For example, in wage gender gap analysis, where Y is wage, D is the gender indicator, and W are various other variables explaining variation in wages. In program evaluation, D is often a treatment or policy variable and W are controls. Write

$$Y = D'\beta_1 + W'\beta_2 + \varepsilon. \tag{3.1}$$

What does the regression coefficient β_1 measure here? It measures how our linear prediction of Y changes if we set the gender variable D from 0 to 1, holding the controls W fixed. We can call this the *predictive effect* (PE), as it measures the impact of a variable on the prediction we make. PE is a measure of statistical dependence or association between variables suggesting that D predicts Y even if we partial-out linearly the controls W. The PE should not be in general interpreted as a causal or treatment effect (TE), since correlation is not equivalent to causation. We shall study assumption needed for causal interpretability of the estimates later in the course. An important case where β_1 measures TE is the case of randomizes control trials, where D is randomly assigned, and is therefore independent of X.

In *population*, define the partialling-out operator with respect to a vector W that takes a random variable V such that $\mathrm{E}V^2 < \infty$ and creates $\tilde{V}$ according to the rule:

$$\tilde{V} = V - W'\gamma_{VW}, \quad \gamma_{VW} = \arg\min_{b \in \mathbb{R}^{p_2}} \mathrm{E}(V - W'b)^2.$$

When V is a vector, we interpret the application of the operator as componentwise. The vector W needs to have finite second moment in order for this to be well-defined.

It is not difficult to see that the partialling-out operator is linear on the space of random variables with finite second moments, i.e. if for V and U such that $\mathrm{E}U^2 + \mathrm{E}V^2 < \infty$,

$$Y = V + U \implies \tilde{Y} = \tilde{V} + \tilde{U}.$$

Thus we apply this operator to both sides of the identity (3.1) to get:

$$\tilde{Y} = \tilde{D}'\beta_1 + \tilde{W}'\beta_2 + \tilde{\varepsilon},$$

which implies that

$$\tilde{Y} = \tilde{D}'\beta_1 + \varepsilon, \quad \mathrm{E}\varepsilon\tilde{D} = 0. \tag{3.2}$$

The last line follows from $\tilde{W} = 0$, which holds by definition, and $\tilde{\varepsilon} = \varepsilon$, which holds because of the orthogonality $\mathrm{E}\varepsilon X = 0$; moreover, since $\tilde{D}$ is a linear combination of components of X, we have that $\mathrm{E}\varepsilon\tilde{D} = 0$.

Equation (3.2) states that $\mathrm{E}\varepsilon\tilde{D} = 0$ is the first-order condition for the population regression of $\tilde{Y}$ on $\tilde{D}$. That is, the projection coefficient β_1 can be recovered from the regression of $\tilde{Y}$ on $\tilde{D}$:

$$\beta_1 = \arg\min_{b\in\mathbb{R}^{p_1}} \mathrm{E}(\tilde{Y} - \tilde{D}'b) = (\mathrm{E}\tilde{D}\tilde{D}')^{-1}\mathrm{E}\tilde{D}\tilde{Y}.$$

This is a remarkable fact, known as Frisch-Waugh-Lovell (FWL) theorem. It asserts that β_1 is a regression coefficient of Y on D after partialing-out the linear effect of W from Y and D. In other words, it measures linearly the predictive effect (PE) of D on Y, after taking out the linear predictive effect of W on both of these variables.

In the *sample*, partialling-out operation works similarly. Define it as an operator that converts V_i into $\check{V}_i$ via

$$\check{V}_i = V_i - W_i'\hat{\gamma}_{VW}, \quad \hat{\gamma}_{VW} = \arg\min_{b\in\mathbb{R}^{p_2}} \mathbb{E}_n(V - W'b)^2.$$

Similarly to the population case, the operator is linear. Thus, application of the operator to the decomposition identity $Y_i \equiv D_i'\hat{\beta}_1 + W_i'\hat{\beta}_2 + \hat{\varepsilon}_i$ gives

$$\check{Y}_i = \check{D}_i'\hat{\beta}_1 + \hat{\varepsilon}_i, \quad \mathbb{E}_n\hat{\varepsilon}\check{D} = 0.$$

This implies that

$$\hat{\beta}_1 = \arg\min_{b\in\mathbb{R}^{p_1}} \mathbb{E}_n(\check{Y} - \check{D}'b) = (\mathbb{E}_n\check{D}\check{D}')^{-1}\mathbb{E}_n\check{D}\check{Y}.$$

This is the sample version of the FWL Theorem.

The partialling-out operation defined above works well when the dimension of W is low in relation to the sample size. When the dimension is high we need to use variable selection or penalization for regularization purposes. We shall get to that later in the course.

We summarize the discussion as a theorem.

Theorem 1 (Frisch-Waugh-Lovell). *Work with the set-up above. The population projection coefficient $\hat{\beta}_1$ can be recovered from the population regression of $\tilde{Y}$ on $\tilde{D}$:*

$$\beta_1 = (\mathrm{E}\tilde{D}\tilde{D}')^{-1}\mathrm{E}\tilde{D}\tilde{Y},$$

assuming $\mathrm{E}\tilde{D}\tilde{D}'$ is of full rank. The sample projection coefficient $\hat{\beta}_1$ can be recovered from the sample regression of $\check{Y}_i$ on $\check{D}_i$:

$$\hat{\beta}_1 = (\mathbb{E}_n\check{D}\check{D}')^{-1}\mathbb{E}_n\check{D}\check{Y},$$

assuming $\mathbb{E}_n\check{D}\check{D}'$ is of full rank.

4. Approximate Distributions for $\hat{\beta}_1$

It is of interest to examine the behavior of the estimator $\hat{\beta}_1$. In what follows, we can assume that dimension p_1 of the target parameter β_1 is fixed, but the dimension p_2 of the nuisance parameter β_2 may grow with n but slowly enough so that $p_2/n \to 0$. In practical terms, the latter condition simply means that p_2 is small compared to n.

Lemma 1 (Adaptivity Property for Partialling Out). *Consider the sample projection coefficient $\hat{\beta}_1$ obtained from the sample regression of $\check{Y}_i$ on $\check{D}_i$:*

$$\hat{\beta}_1 = (\mathbb{E}_n\check{D}\check{D}')^{-1}\mathbb{E}_n\check{D}\check{Y},$$

and the sample projection coefficient $\tilde{\beta}_1$ obtained from the sample regression of infeasible $\tilde{Y}_i$ on $\tilde{D}_i$:

$$\tilde{\beta}_1 = (\mathbb{E}_n\tilde{D}\tilde{D}')^{-1}\mathbb{E}_n\tilde{D}\tilde{Y}.$$

There exist regularity conditions such that, provided that the dimension p_2 is small compared to n, namely

$$p_2/n \to 0,$$

we have the following asymptotic equivalence result:

$$\sqrt{n}(\hat{\beta}_1 - \tilde{\beta}_1) \to_P 0.$$

That is, the estimator is not affected by the estimation errors in partialling out steps, and they are approximately negligible.

We have that

$$\begin{aligned}\tilde{\beta}_1 - \beta_1 &= (\mathbb{E}_n \tilde{D}\tilde{D}')^{-1}\mathbb{E}_n \tilde{D}\tilde{Y} - \beta_1 && (4.1)\\ &= (\mathbb{E}_n \tilde{D}\tilde{D}')^{-1}\mathbb{E}_n \tilde{D}(\beta_1 \tilde{D} + \epsilon) - \beta_1 && (4.2)\\ &= (\mathbb{E}_n \tilde{D}\tilde{D}')^{-1}\mathbb{E}_n \tilde{D}\epsilon. && (4.3)\end{aligned}$$

Then we conclude that under mild regularity conditions

$$\sqrt{n}(\tilde{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, V_{11})$$

where $\overset{a}{\sim}$ reads as "approximately distributed",

$$V_{11} = (\mathrm{E}\tilde{D}\tilde{D}')^{-1} \operatorname{Var}(\sqrt{n}\mathbb{E}_n \tilde{D}\epsilon)(\mathrm{E}\tilde{D}\tilde{D}')^{-1}.$$

Given the equivalence stated in Lemma above we further conclude that

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, V_{11}).$$

Theorem 2. *There exist regularity conditions such that, provided that $p_2/n \to 0$, we have that*

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, V_{11}),$$

as $n \to \infty$, namely that

$$\sup_{A \in \mathcal{A}} \left| \mathrm{P}\left(\sqrt{n}(\hat{\beta}_1 - \beta_1) \in A\right) - \mathrm{P}(N(0, V_{11}) \in A) \right| \to 0,$$

where $\mathcal{A}$ is a collection of sets in $\mathbb{R}^{p_1}$ (e.g. convex sets or rectangles).

The proof of this result is simple under fixed p_2 and is rather technical when $p_2 \to \infty$, so we won't pursue it here, but conceptually it is a more technical version of the result under fixed p asymptotics that you have seen in the introductory regression course.

Remark 1. Alternatively, the result above could also be derived or conjectured from the statement that the whole parameter vector is approximately normally distributed as follows:

$$\sqrt{n}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, V), \tag{4.4}$$

Here

$$V = Q^{-1}\Omega Q^{-1}, \quad Q = \mathrm{E}XX', \quad \Omega = \operatorname{Var}(\sqrt{n}\mathbb{E}_n X\varepsilon).$$

Then V_{11} corresponds to the $p_1 \times p_1$ upper-left block of V. This result is straightforward when p is fixed as $n \to \infty$. On the other hand, when p is increasing with n, proving that the whole p-dimensional parameter vector $\sqrt{n}(\hat{\beta} - \beta)$ is normally distributed is usually much more demanding, in terms of regularity conditions and the sense in which normal approximations hold.

We shall rely on a suitable estimator $\hat{V}_{11}$ of V_{11}, for example, the White estimator under independent sampling or the Newey-West estimator for the time series case. We then shall use the normal law $N(0, \hat{V}_{11}/n)$ for *quantification of uncertainty* about β_1, that is, for building confidence bands for β_1 and various functionals of β. With $\hat{V}_{11}$ used instead of V_{11} the statement $\sqrt{n}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, \hat{V}_{11})$ is defined to mean the following:

$$\sup_{A \in \mathcal{A}} \left| \mathrm{P}\left(\sqrt{n}(\hat{\beta} - \beta) \in A \right) - \mathrm{P}(N(0, \bar{V}) \in A) \mid_{\bar{V} = \hat{V}_{11}} \right| \to_P 0.$$

Basically, we just insert $\hat{V}$ wherever V previously appeared and we require the same statements to hold stochastically.

Lemma 2 (Using Estimated Variance is Ok.). *Suppose that $\sqrt{n}(\hat{\beta}_1 - \beta_1) \overset{a}{\sim} N(0, V_{11})$ and $\hat{V}_{11}$ is consistent for V_{11}, namely $\hat{V}_{11}^{-1} V_{11} \to_P I$ and V_{11} is bounded away from zero. Then $\sqrt{n}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, \hat{V}_{11})$.*

This lemma is a consequence of the Gaussian vector $N(0, V_{11})$ having bounded density, so that estimation errors in $\hat{V}_{11}$ have a negligible effect on probabilities of the containment events.

Suppose β_1 is scalar or we are interested in the j-th component of β_j. The above results means that we can report $\sqrt{\hat{V}_{11,jj}/n}$ as (estimated) standard errors for β_{ij}, and report

$$[\ell_j, u_j] = \left[\hat{\beta}_{1j} - z\sqrt{\hat{V}_{11,jj}/n}, \quad \hat{\beta}_{1j} + z\sqrt{\hat{V}_{11,jj}/n} \right],$$

where z is $(1 - \alpha/2)$-quantile of the standard normal variable $N(0, 1)$, as the approximate $(1 - \alpha) \times 100\%$ confidence interval for β_{1j}. That this is a confidence interval follows from a more general result we discuss below.

5. Gender Wage Gap in 2015

We consider an empirical application to gender wage gap using data from the U.S. March Supplement of the Current Population Survey (CPU) in 2015. We select white non-hispanic individuals, aged 25 to 64 years, and working more than 35 hours per week during at least 50 weeks of the year. We exclude self-employed workers; individuals living in group quarters; individuals in the military, agricultural or private household sectors; individuals with inconsistent reports on earnings and employment status; individuals with allocated or missing information in any of the variables used in the analysis; and individuals with hourly wage below \$3.[1] The resulting sample consists of $32,523$ workers including $18,137$ men and $14,386$ of women. The variable of interest Y is the logarithm of the hourly

[1] The sample selection criteria is similar to [5].

wage rate constructed as the ratio of the annual earnings to the total number of hours worked, which is constructed in turn as the product of number of weeks worked and the usual number of hours worked per week. Table 1 reports descriptive statistics for the variables used in the analysis. Working women are less likely to be married and more highly educated than working men, but have slightly less experience. The unconditional average gender wage gap is 24%.

TABLE 1. Descriptive Statistics

	All	Men	Women
log wage	3.16	3.26	3.02
female	0.44	0.00	1.00
married	0.70	0.73	0.65
widowed	0.01	0.01	0.02
separated	0.02	0.01	0.02
divorced	0.12	0.09	0.15
never married	0.16	0.16	0.16
lhs	0.02	0.03	0.01
hsg	0.25	0.28	0.21
sc	0.28	0.27	0.30
cg	0.28	0.27	0.29
ad	0.17	0.15	0.19
ne	0.19	0.19	0.20
mw	0.26	0.26	0.26
so	0.33	0.33	0.34
we	0.22	0.23	0.21
experience	21.21	21.35	21.03

Source: March Supplement CPS 2015

To estimate the gender wage gap, we consider the linear regression model:

$$Y = D\beta_1 + W'\beta_2 + \varepsilon, \quad \mathrm{E}\varepsilon(D, W')' = 0,$$

where Y is the log hourly rate, D is an indicator for female worker, and W is a set of $p = 1,082$ controls including 5 marital status indicators (widowed, divorced, separated, never married, and married); 5 educational attainment indicators (less than high school graduates, high school graduates, some college, college graduate, and advanced degree); 4 region indicators (midwest, south, west, and northeast); a quartic in potential experience constructed as the maximum of age minus years of schooling minus 7 and zero, i.e.,

$experience = \max(age - education - 7, 0)$; 22 occupation indicators;[2] 21 industry indicators;[3] and all the two-way interactions between the previous variables.

Table 2 reports the results of a regression analysis using the CPS data. The first row obatins the coefficient of D from the OLS regression of Y on D; the second row obtains the coefficient of D from the OLS regression of Y on $X = (D, W)$; the second row obtains the same estimate using the Frisch-Waugh-Lovell theorem for partialing-out the controls via OLS; and the third row obtains the coefficient of D using a variant of the procedure in [1] that partials-out the controls via LASSO instead of OLS.[4] We will study this procedure later in the course. All the standard errors are computed with the `R` package `sandwich` and are robust to heteroskedasticity. Using Lasso for partialing out here gives similar results as using OLS. Lasso is a penalized OLS estimator and it produces high-quality estimates of the regression function especially in the high-dimensional settings. The penalty takes the form of the sum of the absolute values of the coefficients times penalty level.

TABLE 2. Regression Analysis of the Wage Gap

	Estimate	Std. Error†
no controls	-0.239	0.0067
all controls	-0.185	0.0069
partial reg	-0.185	0.0069
partial reg with lasso	-0.195	0.0068

† Standard errors are robust to heteroskedasticity

What do the estimated regression coefficients β_1 measure here? The first row measures the unconditional gender gap, i.e. the difference in the average wage of working women and men. The rest measure how our linear prediction of wage changes if we set the gender variable D from 0 to 1, holding the controls W fixed. We can call this the *predictive effect* (PE), as it measures the impact of a variable on the prediction we make. The PE should

[2]The occupation categories are: management; business and financial operations; computer and mathematics; architecture and engineering; life, physical, and social science; community and social service; legal; education, training, and library; arts, design, entertainment, sports, and media; healthcare practitioners and technical; healthcare support; protective service; food preparation and serving; building and grounds cleaning and maintenance; personal care and service; sales; office and administrative support; farming, fishing, and forestry; construction and extraction; installation, maintenance, and repair occupations; production; and transportation and material moving.

[3]The industry categories are: mining; utilities; construction; nondurable goods manufacturing; durable goods manufacturing; durable goods wholesale; nondurable goods wholesale; retail trade; transportation and warehousing; information; finance and insurance; real estate, rental and leasing; professional, scientific, and technical services; management of companies and enterprises; administrative, support and waste management services; educational services; health care and social assistance; arts, entertainment, and recreation; accommodation and food services; other services except public administration; and public administration.

[4]We use the `R` package `hdm` to obtain the estimates in the third row.

not be in general interpreted as a causal or treatment effect (TE), since correlation is not equivalent to causation. The causal interpretation of PE here could suggest that β_1 is solely a measure of discrimination, while in reality it may reflect discrimination, selection effects (e.g., sorting of women and men into different occupation), sample imbalances, etc. In this case the unconditional wage gap for women of 24% decreases to around 19-20% after controlling for worker characteristics.

We repeat the analysis for the more homogeneous subpopulation of never married workers. Table 3 reports descriptive statistics for the corresponding subsample from the CPS 2015 data. There are 5,150 never married workers, 2,861 men and 2,289 women. Never married working women are also relatively more educated than working men. Compared to Table 1, never married workers earn lower average wages, and have much lower experience than the rest of the workers. The regression analysis in Table 4 shows that the unconditional gender wage gap is less than 4% for this group. This gap increases to 6-7% once we control for worker characteristics.[5] A possible explanation of the lower wage gap for never married working women could be related to fertility and childcare decisions. Thus, never married women are young and less likely to have children. They can therefore be more career oriented and have working experiences not interrupted by childbearing or childcare.

TABLE 3. Descriptive Statistics: Never Married Workers

	All	Male	Female
log wage	2.97	2.99	2.95
female	0.44	0.00	1.00
lhs	0.02	0.03	0.01
hsg	0.24	0.29	0.18
sc	0.28	0.27	0.28
cg	0.32	0.29	0.35
ad	0.14	0.11	0.18
ne	0.23	0.22	0.24
mw	0.26	0.26	0.26
so	0.30	0.30	0.29
we	0.22	0.22	0.21
experience	13.76	13.78	13.73

Source: March Supplement CPS 2015

[5] Without the marital status indicators, there are $p = 775$ controls.

TABLE 4. Regression Analysis of the Wage Gap: Never Married Workers

	Estimate	Std. Error†
No controls	-0.038	0.016
All controls	-0.061	0.015
partial reg	-0.061	0.015
partial reg via lasso	-0.070	0.015

† Standard errors are robust to heteroskedasticity

6. Joint Confidence Bands for β_1

Consider a p_1-dimensional subvector β_1 of the coefficient vector β. Assume, without loss of generality, that these are the first p_1 components. Assume that

$$\hat{\beta}_1 - \beta_1 \overset{a}{\sim} N(0, V_{11}/n),$$

where V_{11} is the upper-left $p_1 \times p_1$ sub-block of V, in the sense that

$$\sup_{A \in \mathcal{A}} \left| \mathrm{P}\left(\sqrt{n}(\hat{\beta}_1 - \beta_1) \in A\right) - \mathrm{P}(N(0, V_{11}) \in A) \right| \to 0, \quad n \to \infty, \tag{6.1}$$

where $\mathcal{A}$ is a collection of rectangles in $\mathbb{R}^{p_1}$.

Suppose we want to build simultaneous confidence bands for all the components $(\beta_{1j})_{j=1}^{p_1}$ of β_1. To give a context, suppose that D represents a collection of indicator ("dummy") variables, capturing different types of treatment. For instance, in the Pennsylvania treatment experiments example below, the components of D describe various kinds of incentives that participants received to find a job quicker. We want to create a confidence set $[\ell, u] = ([\ell_j, u_j])_{j=1}^{p_1}$ such that

$$\mathrm{P}(\beta_1 \in [\ell, u]) = \mathrm{P}\left(\beta_{1j} \in [\ell_j, u_j] \text{ for all } j\right) \to 1 - \alpha.$$

Using such confidence bands allows us to answer a number of very interesting questions about both economic and statistical significance of the components of β_1. For example, we can create a confidence set for a set of treatments that result in more than some target level of impact.

We consider confidence bands in the form of the rectangle:

$$[\ell_j, u_j] = \left[\hat{\beta}_{1j} - c\sqrt{V_{11,jj}/n}, \quad \hat{\beta}_{1j} + c\sqrt{V_{11,jj}/n} \right], \quad j = 1, ..., p_1,$$

where we set the critical value c such that the previous display holds. Here we use $V_{11,jj}$ to denote the (j, j)-th element of matrix V_{11}.

The value of c can be determined as $(1-\alpha)$-quantile of

$$\|N(0,C)\|_\infty,$$

where C is the correlation matrix associated with V_{11}, that is,

$$C = S^{-1/2} V_{11} S^{-1/2}$$

where $S = \mathrm{diag}(V_{11})$ is a diagonal matrix with the diagonal of V_{11} in its diagonal and zeroes elsewhere. The constant c can be approximated by simulation.

This constant c is the right one by the following argument:

$$\begin{aligned}
\mathrm{P}(\beta_1 \in [\ell, u]) \quad &= \mathrm{P}(\sqrt{n}(\hat{\beta}_1 - \beta_1) \in S^{1/2}[-c, c]) \\
&= \mathrm{P}(N(0, V_{11}) \in S^{1/2}[-c, c]) + o(1) \\
&= \mathrm{P}(S^{-1/2} N(0, V_{11}) \in [-c, c]) + o(1) \\
&= \mathrm{P}(\|N(0, S^{-1/2} V_{11} S^{-1/2})\|_\infty \leq c) + o(1) \\
&= 1 - \alpha + o(1),
\end{aligned}$$

where the second equality holds by (6.1), because $S^{1/2}[-c, c]$ is a rectangular set in $\mathbb{R}^{p_1}$.

Note that in practice we shall need to replace V_{11} with a consistent estimator $\hat{V}_{11}$. This replacement does not affect the approximate coverage property of the confidence regions in view of Lemma 2.

We summarize the discussion as follows.

Theorem 3 (Joint Confidence Band For Target Coefficients). *Suppose that* $\hat{\beta}_1 - \beta_1 \overset{\mathrm{a}}{\sim} N(0, V_{11}/n)$ *in the sense of (6.1). We have that the confidence band*

$$[\ell_j, u_j] = \left[\hat{\beta}_{1j} - c\sqrt{V_{11,jj}/n}, \quad \hat{\beta}_{1j} + c\sqrt{V_{11,jj}/n} \right], \quad j = 1, ..., p_1,$$

with $c = (1-\alpha)$*-quantile of* $\|N(0, C\|_\infty$*, where* C *is the correlation matrix associated to* V_{11}*, jointly covers all target parameter values* $(\beta_{1j})_{j=1}^{p_1}$ *with probability approaching the nominal level, that is, as* $n \to \infty$*,*

$$\mathrm{P}\left(\beta_{1j} \in [\ell_j, u_j] \text{ for all } j\right) \to 1 - \alpha.$$

The results continue to hold if V_{11} *is replaced by* $\hat{V}_{11}$*, such that* $\hat{V}_{11}^{-1} V_{11} \to_P I$ *and* V_{11} *is bounded away from zero.*

Here we re-analyze the Pennsylvania re-employment bonus experiment, which was previously studied in [2], among others. Note that the inferential results on simultaneous bands we report below will be new. These experiments were conducted in the 1980s by the U.S. Department of Labor to test the incentive effects of alternative compensation schemes for unemployment insurance (UI). In these experiments, UI claimants were randomly assigned either to a control group or one of five treatment groups.[6] In the control group the current rules of the UI applied. Individuals in the treatment groups were offered a cash bonus if they found a job within some pre-specified period of time (qualification period), provided that the job was retained for a specified duration. The treatments differed in the level of the bonus, the length of the qualification period, and whether the bonus was declining over time in the qualification period; see [2] for further details on data.

To evaluate the impact of the treatments on unemployment duration, we consider the linear regression model:

$$Y = D'\beta_1 + W'\beta_2 + \varepsilon, \quad \mathrm{E}\varepsilon(D', W')' = 0,$$

where Y is the log of duration of unemployment, D is a vector of 5 treatment indicators, and W is a set of $p = 16$ controls including age group dummies, gender, race, number of dependents, quarter of the experiment, location within the state, existence of recall expectations, and type of occupation.

The assignment of units to treatment D is *random*. We commonly refer to such case as the *randomized control trial* (RCT). Under RCT, the projection coefficient β_1 has the interpretation of the causal effect of the treatment on the average outcome. We thus refer to β_1 as the *average treatment effect* (ATE). Note that covariates W here are independent of the treatment D, so we can identify β_1 by just regression of Y on D, without adding covariates. However we do add covariates in an effort to improve the precision of our estimates of the average treatment effect.

Figure Figure 7 shows 90% confidence intervals for the five treatment effects β_1, constructed using a sample of 13,913 observations.

- The critical value for the simultaneous bands, $c = 2.27$, is greater than the pointwise critical value, 1.65.
- It is less than the critical value from the Bonferroni correction, 2.33, obtained as the $(1 - \bar{\alpha}/2)$ quantile of the normal distribution with $\bar{\alpha} = \alpha/5$. The idea of Bonferroni

[6] There are six treatment groups in the experiments. Following [2] we merge the groups 4 and 6.

correction is to use the union bound $\mathrm{P}(\cup_{j=1}^{p_1}\mathrm{event}_j) \leq \sum_{j=1}^{p_1}\mathrm{P}(\mathrm{event}_j)$ to bound the noncoverage event, i.e. $\mathrm{event}_j = \{\beta_j \notin [\ell_j, u_j]\}$.

In this case, from the three treatment levels with statistically significant effect on unemployment duration based on pointwise confidence intervals, only one remains significant after accounting for simultaneous inference.

The last observation illustrates how econometrics and this class offer better concepts and tools than what the standard empirical practice often does. It also explains why econometricians sometimes teach what "people never use in practice" – they simply teach correct things to use, and it is up to you to decide whether you want to do wrong or correct things in practice.

Next we consider a more flexible version of the more basic model, where we take controls to include the original set set of controls as well as all two-way interactions, giving us a total of $p = 120$ controls. We repeat the exercise we have given above with roughly similar conclusions. Figure 7 shows 90% confidence intervals for the five treatment effects β_1. We see that the addition of many more controls does not change the inferential results noticeably. This highlights the robustness of the conclusions with respect to enriching the set of controls, and is also in-line with our asymptotic theory, which states that the inference is not impacted in the regime where the number of controls p is much smaller than n, despite the fact the number of controls is substantial here.

Notes

Least squares were invented by Legendre around 1800, although Gauss claimed the credit. Frisch, Waugh, and Lovell discovered the partialling out interpretation of the least squares coefficients in 1930s. The adaptivity results of Lemma 1 went unnoticed by empiricists, and also manage to escape statistics and econometrics textbooks; we note this property here though. Regularity conditions under which Lemma 1 and Theorem 2 hold under fixed p asymptotitcs can be found in the introductory econometrics texts, for example, [6], and under $p \to \infty$ and $p/n \to 0$ asymptotitcs in [4] and [3]. The results of the latter reference allow for $p/n \to c$, which introduces an additional asymptotic variance term, and the case with $c = 0$ recovers Theorem 2.

Problems

(1) Briefly explain partialling-out and the adaptitive property for the linear regression model, and use the gender wage gap data to illustrate your points. Present your

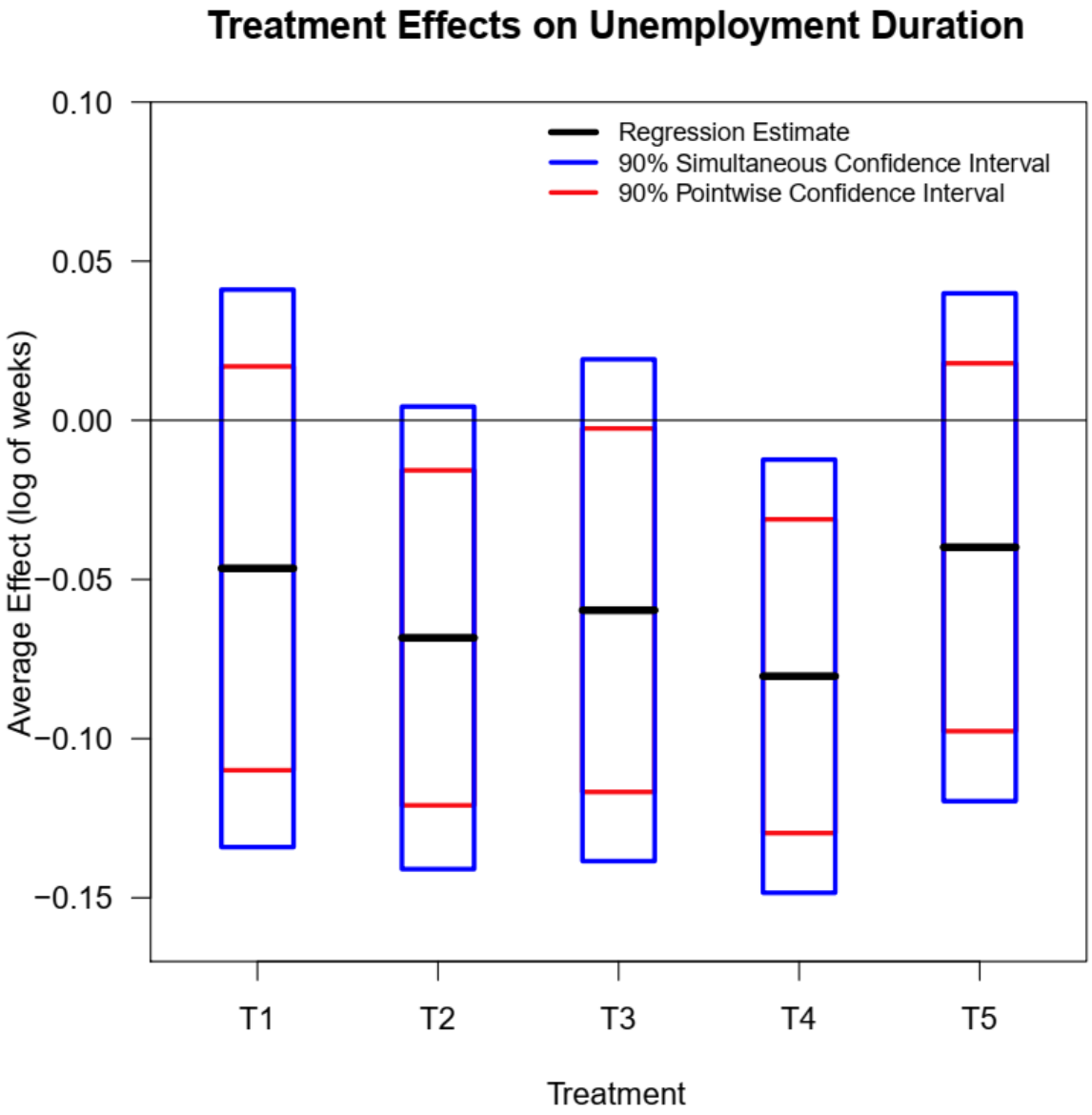

FIGURE 1. 90% Confidence Intervals for Treatment Effects on Unemployment Duration. Number of controls is 16. Critical value for simultaneous confidence interval obtained by simulation with 100,000 repetitions.

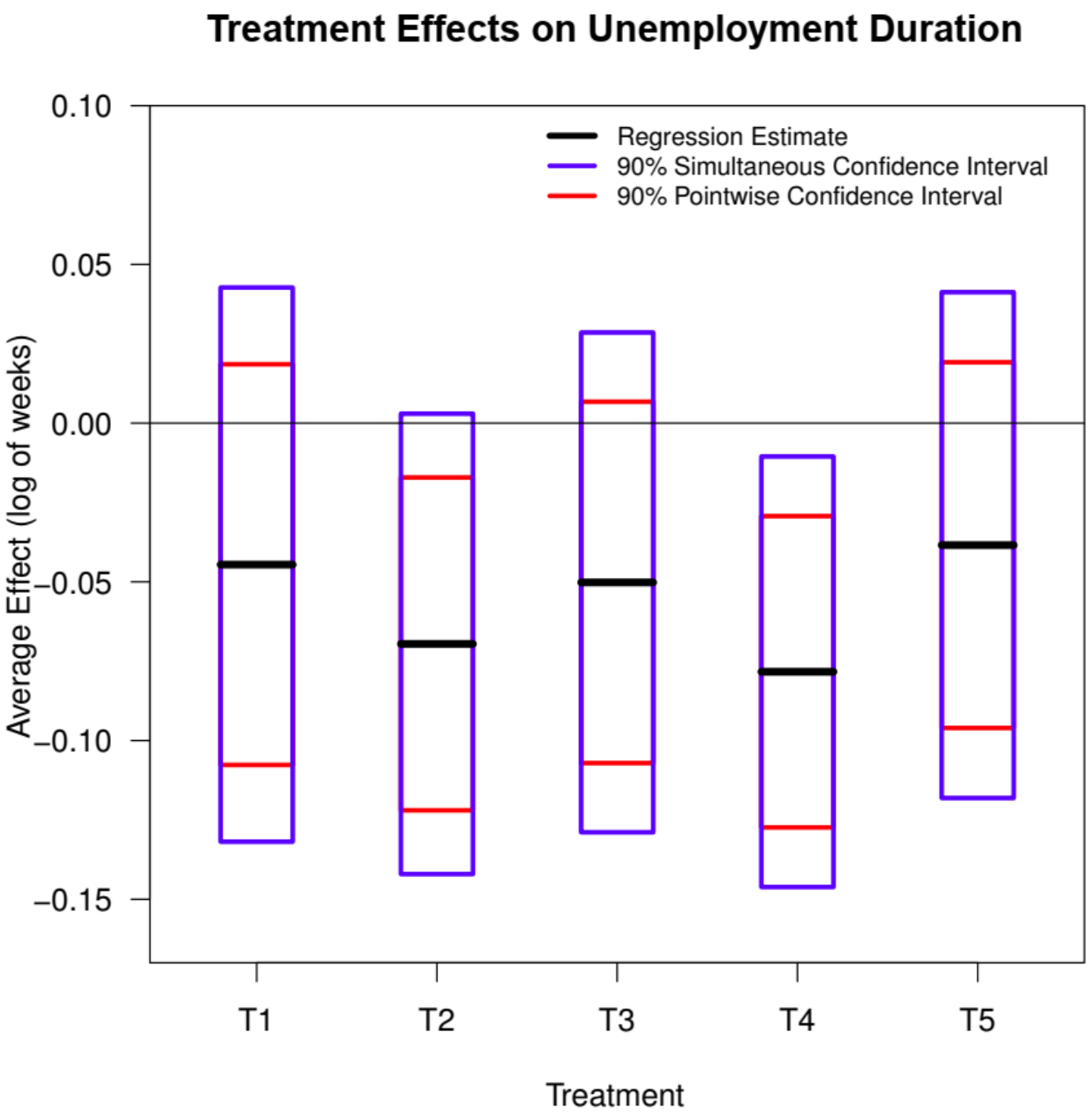

FIGURE 2. 90% Confidence Intervals for Treatment Effects on Unemployment Duration. Number of controls is 120. Critical value for simultaneous confidence interval obtained by simulation with 100,000 repetitions.

discussion as a brief section of a professionally done empirical paper.

(2) Briefly explain the idea of joint confidence bands, and use the Penn data to replicate the second set of results of our re-analysis of Pennsylvania re-employment experiment. Present your results as a brief section of a professionally done empirical paper.

(3) In the wage gap example and reemployment experiment, discuss whether the empirical results have a causal or treatment effect interpretation. Does the estimate wage gap measure discrimination? Perhaps in part? Do the reductions in unemployment duration have a causal meaning? Present your discussion as a brief section of a professionally done empirical paper.

(4) Explain why in randomized control trials, where assigned treatment D is independent from controls W, we can estimate the linear predictive effect of D on Y controlling linearly for W without actually controlling for W. However, including W may still be a good idea, because using W can lower (and does not increase) the asymptotic variance of the least squares estimator.

(5) Prove that the population partialling-out operator is linear on the space of random variables with finite second moments, i.e. if for V and U such that $\mathrm{E}U^2 + \mathrm{E}V^2 < \infty$,

$$Y = V + U \implies \tilde{Y} = \tilde{V} + \tilde{U}.$$

(6) Provide a set of sufficient regularity conditions for Lemma 1 and Theorem 1 and prove them. Extra credit is given for handling the case where $p_2 \to \infty$ as $n \to \infty$, but don't spend too much time on this, as this is difficult.

References

[1] Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

[2] Bilias, Y. (2000): "Sequential testing of duration data: the case of the Pennsylvania Reemployment bonus experiment," *Journal of Applied Econometrics*, 15(6), 575–594.

[3] Cattaneo, M. D., M. Jansson, and W. K. Newey (2015): "Inference in Linear Regression Models with Many Covariates and Heteroskedasticity," *ArXiv e-prints*.

[4] Chernozhukov, V., D. Chetverikov, and K. Kato (2015): "Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results," *Journal of Econometrics*, 186, 345–366.